Computer Scanner and Antiviurs Programmes

Neil Reynolds

COMPUTER SCANNER AND ANTIVIURS PROGRAMMES

COMPUTER SCANNER AND ANTIVIURS PROGRAMMES

Neil Reynolds

BIBLIOTEX
Digital Library

Computer Scanner and Antiviurs Programmes
by Neil Reynolds

www.bibliotex.com

Ebook ISBN: 9781984664266

Published by:
Bibliotex
Canada
Website: www.bibliotex.com

Contents

1

3D Scanner

A 3D scanner is a device that analyzes a real-world object or environment to collect data on its shape and possibly its appearance (i.e. color). The collected data can then be used to construct digital, three dimensional models. Many different technologies can be used to build these 3D scanning devices; each technology comes with its own limitations, advantages and costs. Many limitations in the kind of objects that can be digitized are still present: for example optical technologies encounter many difficulties with shiny, mirroring or transparent objects. Collected 3D data is useful for a wide variety of applications. These devices are used extensively by the entertainment industry in the production of movies and video games. Other common applications of this technology include industrial design, orthotics and prosthetics, reverse engineering and prototyping, quality control/inspection and documentation of cultural artifacts.

Functionality

The purpose of a 3D scanner is usually to create a point cloud of geometric samples on the surface of the subject. These points can then be used to extrapolate the shape of the subject (a process called reconstruction). If color information is collected at each point, then the colors on the surface of the subject can also be determined. 3D scanners are very analogous to cameras. Like cameras, they have a cone-like field of view, and like cameras, they can only collect information about surfaces that are not obscured. While a camera collects color information about surfaces within its field of view, 3D scanners collect distance information about surfaces within its field of view.

The "picture" produced by a 3D scanner describes the distance to a surface at each point in the picture. If a spherical coordinate system is defined in which the scanner is the origin and the vector out from the front of the scanner is ö=0 and è=0, then each point in the picture is associated with a ö and è. Together with distance, which corresponds to the r component, these spherical coordinates fully describe the three dimensional position of each point in the picture, in a local coordinate system relative to the scanner.

For most situations, a single scan will not produce a complete model of the subject. Multiple scans, even hundreds, from many different directions are usually required to obtain information about all sides of the subject. These scans have to be brought in a common reference system, a process that is usually called *alignment* or *registration*, and then merged to create a complete model.

This whole process, going from the single range map to the whole model, is usually known as the 3D scanning pipeline.

Technology

There are a variety of technologies for digitally acquiring the shape of a 3D object. A well established classification divides them into two types: contact and non-contact 3D scanners. Non-contact 3D scanners can be further divided into two main categories, active scanners and passive scanners. There are a variety of technologies that fall under each of these categories.

Contact

Contact 3D scanners probe the subject through physical touch. A CMM (coordinate measuring machine) is an example of a contact 3D scanner. It is used mostly in manufacturing and can be very precise. The disadvantage of CMMs though, is that it requires contact with the object being scanned. Thus, the act of scanning the object might modify or damage it.

This fact is very significant when scanning delicate or valuable objects such as historical artifacts. The other disadvantage of CMMs is that they are relatively slow compared to the other scanning methods. Physically moving the arm that the probe is mounted on can be very slow and the fastest CMMs can only operate on a few hundred hertz. In contrast, an optical system like a laser scanner can operate from 10 to 500 kHz. Other examples are the hand driven touch probes used to digitize clay models in computer animation industry.

Non-contact Active

Active scanners emit some kind of radiation or light and detect its reflection in order to probe an object or environment. Possible types of emissions used include light, ultrasound or x-ray.

Time-of-flight

The time-of-flight 3D laser scanner is an active scanner that uses laser light to probe the subject. At the heart of this type of scanner is a time-of-flight laser rangefinder. The laser rangefinder finds the distance of a surface by timing the round-trip time of a pulse of light. A laser is used to emit a pulse of light and the amount of time before the reflected light is seen by a detector is timed.

Since the speed of light *c* is known, the round-trip time determines the travel distance of the light, which is twice the distance between the scanner and the surface. If *t* is the round-trip time, then distance is equal to . The accuracy of a time-of-flight 3D laser scanner depends on how precisely we can measure the *t* time: 3.3 picoseconds (approx.) is the time taken for light to travel 1 millimetre.

The laser rangefinder only detects the distance of one point in its direction of view. Thus, the scanner scans its entire field of view one point at a time by changing the range finder's direction of view to scan different points. The view direction of the laser rangefinder can be changed either by rotating the range finder itself, or by using a system of rotating mirrors.

The latter method is commonly used because mirrors are much lighter and can thus be rotated much faster and with

greater accuracy. Typical time-of-flight 3D laser scanners can measure the distance of 10,000~100,000 points every second. Time-of-flight devices are also available in a 2D configuration. This is referred to as a Time-of-flight camera.

Triangulation

The triangulation 3D laser scanner is also an active scanner that uses laser light to probe the environment. With respect to time-of-flight 3D laser scanner the triangulation laser shines a laser on the subject and exploits a camera to look for the location of the laser dot.

Depending on how far away the laser strikes a surface, the laser dot appears at different places in the camera's field of view.

This technique is called triangulation because the laser dot, the camera and the laser emitter form a triangle. The length of one side of the triangle, the distance between the camera and the laser emitter is known. The angle of the laser emitter corner is also known. The angle of the camera corner can be determined by looking at the location of the laser dot in the camera's field of view.

These three pieces of information fully determine the shape and size of the triangle and gives the location of the laser dot corner of the triangle. In most cases a laser stripe, instead of a single laser dot, is swept across the object to speed up the acquisition process.

The National Research Council of Canada was among the first institutes to develop the triangulation based laser scanning technology in 1978.

Notes on Time-of-flight and Triangulation Scanners

Time-of-flight and triangulation range finders each have strengths and weaknesses that make them suitable for different situations. The advantage of time-of-flight range finders is that they are capable of operating over very long distances, on the order of kilometers.

These scanners are thus suitable for scanning large structures like buildings or geographic features. The disadvantage of time-of-flight range finders is their accuracy. Due to the high speed of light, timing the round-trip time is difficult and the accuracy of the distance measurement is relatively low, on the order of millimeters. Triangulation range finders are exactly the opposite. They have a limited range of some meters, but their accuracy is relatively high. The accuracy of triangulation range finders is on the order of tens of micrometers.

Time of flight scanners accuracy can be lost when the laser hits the edge of an object because the information that is sent back to the scanner is from two different locations for one laser pulse. The coordinate relative to the scanners position for a point that has hit the edge of an object will be calculated based on an average and therefore will put the point in the wrong place.

When using a high resolution scan on an object the chances of the beam hitting an edge are increased and the resulting data will show noise just behind the edges of the object. Scanners with a smaller beam width will help to solve this problem but will be limited by range as the beam

width will increase over distance. Software can also help by determining that the first object to be hit by the laser beam should cancel out the second. At a rate of 10,000 sample points per second, low resolution scans can take less than a second, but high resolution scans, requiring millions of samples, can take minutes for some time-of-flight scanners. The problem this creates is distortion from motion. Since each point is sampled at a different time, any motion in the subject or the scanner will distort the collected data. Thus, it is usually necessary to mount both the subject and the scanner on stable platforms and minimize vibration. Using these scanners to scan objects in motion is very difficult.

Recently, there has been research on compensating for distortion from small amounts of vibration. When scanning in one position for any length of time slight movement can occur in the scanner position due to changes in temperature. If the scanner is set on a tripod and there is strong sunlight on one side of the scanner then that side of the tripod will expand and slowly distort the scan data from one side to another. Some laser scanners have level compensators built into them to counteract any movement of the scanner during the scan process.

Conoscopic Holography

In a Conoscopic system, a laser beam is projected onto the surface and then the immediate reflection along the same ray-path are put through a conoscopic crystal and projected onto a CCD. The result is a diffraction pattern, that can be frequency analyzed to determine the distance to the measured surface. The main advantage with

Conoscopic Holography is that only a single ray-path is needed for measuring, thus giving an opportunity to measure for instance the depth of a finely drilled hole.

Hand-held Laser

Hand-held laser scanners like Perceptron's ScanWorks Solutions create a 3D image through the triangulation mechanism described above: a laser dot or line is projected onto an object from a hand-held device and a sensor (typically a charge-coupled device or position sensitive device) measures the distance to the surface. Data is collected in relation to an internal coordinate system and therefore to collect data where the scanner is in motion the position of the scanner must be determined.

The position can be determined by the scanner using reference features on the surface being scanned (typically adhesive reflective tabs) or by using an external tracking method. External tracking often takes the form of a laser tracker (to provide the sensor position) with integrated camera (to determine the orientation of the scanner) or a photogrammetric solution using 3 or more cameras providing the complete Six degrees of freedom of the scanner. Both techniques tend to use infrared Light-emitting diodes attached to the scanner which are seen by the camera(s) through filters providing resilience to ambient lighting.

Data is collected by a computer and recorded as data points within Three-dimensional space, with processing this can be converted into a triangulated mesh and then a Computer-aided design model, often as Nonuniform rational B-spline surfaces. Hand-held laser scanners can combine

this data with passive, visible-light sensors - which capture surface textures and colors - to build (or "reverse engineer") a full 3D model.

Structured Light

Structured-light 3D scanners project a pattern of light on the subject and look at the deformation of the pattern on the subject. The pattern may be one dimensional or two dimensional. An example of a one dimensional pattern is a line. The line is projected onto the subject using either an LCD projector or a sweeping laser. A camera, offset slightly from the pattern projector, looks at the shape of the line and uses a technique similar to triangulation to calculate the distance of every point on the line. In the case of a single-line pattern, the line is swept across the field of view to gather distance information one strip at a time.

An example of a two-dimensional pattern is a grid or a line stripe pattern. A camera is used to look at the deformation of the pattern, and an algorithm is used to calculate the distance at each point in the pattern. Consider an array of parallel vertical laser stripes sweeping horizontally across a target. In the simplest case, one could analyze an image and assume that the left-to-right sequence of stripes reflects the sequence of the lasers in the array, so that the leftmost image stripe is the first laser, the next one is the second laser, and so on. In non-trivial targets having holes, occlusions, and rapid depth changes, however, this sequencing breaks down as stripes are often hidden and may even appear to change order, resulting in laser stripe ambiguity. This problem can be solved using algorithms for

multistripe laser triangulation. Structured-light scanning is still a very active area of research with many research papers published each year. The advantage of structured-light 3D scanners is speed. Instead of scanning one point at a time, structured light scanners scan multiple points or the entire field of view at once. This reduces or eliminates the problem of distortion from motion. Some existing systems are capable of scanning moving objects in real-time.

A real-time scanner a using digital fringe projection and phase-shifting technique (a various structured light method) was developed, to capture, reconstruct, and render high-density details of dynamically deformable objects (such as facial expressions) at 40 frames per second. Recently, another scanner is developed. Different patterns can be applied to this system. The frame rate for capturing and data processing achieves 120 frames per second. It can also scan isolated surfaces, for example two moving hands.

Modulated Light

Modulated light 3D scanners shine a continually changing light at the subject. Usually the light source simply cycles its amplitude in a sinusoidal pattern. A camera detects the reflected light and the amount the pattern is shifted by determines the distance the light traveled. Modulated light also allows the scanner to ignore light from sources other than a laser, so there is no interference.

Volumetric Techniques

Medical: Computed tomography (CT) is a medical imaging method which generates a three-dimensional image of the inside of an object from a large series of two-dimensional

X-ray images, similarly Magnetic resonance imaging is another a medical imaging technique that provides much greater contrast between the different soft tissues of the body than computed tomography (CT) does, making it especially useful in neurological (brain), musculoskeletal, cardiovascular, and oncological (cancer) imaging. These techniques produce a discrete 3D volumetric representation that can be directly visualized, manipulated or converted to traditional 3D surface by mean of isosurface extraction algorithms .

Industrial: Although most common in medicine, Computed tomography, Microtomography and MRI are also used in other fields for acquiring a digital representation of an object and its interior, such as nondestructive materials testing, reverse engineering, or the study biological and paleontological specimens.

Non-contact Passive

Passive scanners do not emit any kind of radiation themselves, but instead rely on detecting reflected ambient radiation. Most scanners of this type detect visible light because it is a readily available ambient radiation. Other types of radiation, such as infrared could also be used. Passive methods can be very cheap, because in most cases they do not need particular hardware but simple digital cameras.

- Stereoscopic systems usually employ two video cameras, slightly apart, looking at the same scene. By analyzing the slight differences between the images seen by each camera, it is possible to determine the

distance at each point in the images. This method is based on the same principles driving human stereoscopic vision.

- Photometric systems usually use a single camera, but take multiple images under varying lighting conditions. These techniques attempt to invert the image formation model in order to recover the surface orientation at each pixel.
- Silhouette techniques use outlines created from a sequence of photographs around a three-dimensional object against a well contrasted background. These silhouettes are extruded and intersected to form the visual hull approximation of the object. With these approaches some concavities of an object (like the interior of a bowl) cannot be detected.

User Assisted

There are other methods that, based on the user assisted detection and identification of some features and shapes on a set of different pictures of an object are able to build an approximation of the object itself.

This kind of techniques are useful to build fast approximation of simple shaped objects like buildings. Various commercial packages are available like iModeller, D-Sculptor or Autodesk ImageModeler. This sort of 3D scanning is based on the principles of photogrammetry.

It is also somewhat similar in methodology to panoramic photography, except that the photos are taken of one object on a three-dimensional space in order to replicate it instead

of taking a series of photos from one point in a three-dimensional space in order to replicate the surrounding environment.

Reconstruction, or Modeling From Point Clouds

The point clouds produced by 3D scanners can be used directly for measurement and visualization in the architecture and construction world. Most applications, however, use instead polygonal 3D models, NURBS surface models, or editable feature-based CAD models (aka Solid models).

Polygon Mesh Models

In a polygonal representation of a shape, a curved surface is modeled as many small faceted flat surfaces (think of a sphere modeled as a disco ball). Polygon models—also called Mesh models, are useful for visualization, for some CAM (i.e., machining), but are generally "heavy" (i.e., very large data sets), and are relatively un-editable in this form. Reconstruction to polygonal model involves finding and connecting adjacent points with straight lines in order to create a continuous surface. Many applications, both free and non free, are available for this purpose (eg. MeshLab, kubit PointCloud for AutoCAD, JRC 3D Reconstructor, photomodeler, imagemodel, PolyWorks, Rapidform, Geomagic, Imageware, Rhino etc.).

Surface Models

The next level of sophistication in modeling involves using a quilt of *curved* surface patches to model our shape.

These might be NURBS, TSplines or other curved representations of curved topology. Using NURBS, our sphere is a true mathematical sphere. Some applications offer patch layout by hand but the best in class offer both automated patch layout and manual layout. These patches have the advantage of being lighter and more manipulable when exported to CAD. Surface models are somewhat editable, but only in a sculptural sense of pushing and pulling to deform the surface. This representation lends itself well to modeling organic and artistic shapes. Providers of surface modelers include Rapidform, Geomagic, Rhino, Maya, T Splines etc.

Solid CAD Models

From an engineering/manufacturing perspective, the ultimate representation of a digitized shape is the editable, parametric CAD model. After all, CAD is the common "language" of industry to describe, edit and maintain the shape of the enterprise's assets. In CAD, our sphere is described by parametric features which are easily edited by changing a value (e.g., centerpoint and radius).

These CAD models describe not simply the envelope or shape of the object, but CAD models also embody the "design intent" (i.e., critical features and their relationship to other features). An example of design intent not evident in the shape alone might be a brake drum's lug bolts, which must be concentric with the hole in the center of the drum. This knowledge would drive the sequence and method of creating the CAD model; a designer with an awareness of this relationship would not design the lug bolts referenced

to the outside diameter, but instead, to the center. A modeler creating a CAD model will want to include both Shape and design intent in the complete CAD model.

Vendors offer different approaches to getting to the parametric CAD model. Some export the NURBS surfaces and leave it to the CAD designer to complete the model in CAD (e.g., Geomagic, Imageware, Rhino). Others use the scan data to create an editable and verifiable feature based model that is imported into CAD with full feature tree intact, yielding a complete, native CAD model, capturing both shape and design intent (e.g. Geomagic, Rapidform). Still other CAD applications are robust enough to manipulate limited points or polygon models within the CAD environment (e.g., Catia).

From a Set of 2D Slices

CT, MRI, or Micro-CT scanners do not produce point clouds but a set of 2D slices (each termed a "tomogram") which are then 'stacked together' to produce a 3D representation. There are several ways to do this depending on the output required:

Volume Rendering

Different parts of an object usually have different threshold values or greyscale densities. From this, a 3-dimensional model can be constructed and displayed on screen. Multiple models can be constructed from various different thresholds, allowing different colors to represent each component of the object. Volume rendering is usually only used for visualisation of the scanned object.

Image Segmentation

Where different structures have similar threshold/ greyscale values, it can become impossible to separate them simply by adjusting volume rendering parameters. The solution is called segmentation, a manual or automatic procedure that can remove the unwanted structures from the image. Image segmentation software usually allows export of the segmented structures in CAD or STL format for further manipulation.

Image-based Meshing

When using 3D image data for computational analysis (e.g. CFD and FEA), simply segmenting the data and meshing from CAD can become time consuming, and virtually intractable for the complex topologies typical of image data. The solution is called image-based meshing, an automated process of generating an accurate and realistic geometrical description of the scan data.

Applications

Material Processing and Production

Laser scanning describes a method where a surface is sampled or scanned using laser technology. Several areas of application exist that mainly differ in the power of the lasers that are used, and in the results of the scanning process.

Lasers with low power are used when the scanned surface doesn't have to be influenced, e.g. when it has to be digitized. Confocal or 3D laser scanning are methods to get information about the scanned surface.

Depending on the power of the laser, its influence on a working piece differs: lower power values are used for laser engraving, where material is partially removed by the laser. With higher powers the material becomes fluid and laser welding can be realized, or if the power is high enough to remove the material completely, then laser cutting can be performed. Also for rapid prototyping a laser scanning procedure is used when for example a prototype is generated by laser sintering.

The principle that is used for all these applications is the same: software that runs on a PC or an embedded system and that controls the complete process is connected with a scanner card. That card converts the received vector data to movement information which is sent to the scanhead. This scanhead consists of two mirrors that are able to deflect the laser beam in one level (X- and Y-coordinate). The third dimension is - if necessary - realized by a specific optic that is able to move the laser's focal point in the depth-direction (Z-axis).

The third dimension is needed for some special applications like the rapid prototyping where an object is built up layer by layer or for in-glass-marking where the laser has to influence the material at specific positions within it. For these cases it is important that the laser has as small a focal point as possible.

For enhanced laser scanning applications and/or high material throughput during production, scanning systems with more than one scanhead are used. Here the software has to control what is done exactly within such a multihead application: it is possible that all available heads have to

mark the same to finish processing faster or that the heads mark one single job in parallel where every scanhead performs a part of the job in case of large working areas. Structured light projection systems are also used for solar cell flatness metrology enabling stress calculation with throughput in excess of 2000 wafers per hour.

Construction Industry and Civil Engineering

- Robotic Control: e.g., a laser scanner may function as the "eye" of a robot.
- As-built drawings of Bridges, Industrial Plants, and Monuments
- Documentation of historical sites
- Site modeling and lay outing
- Quality control
- Quantity Surveys
- Freeway Redesign
- Establishing a bench mark of pre-existing shape/ state in order to detect structural changes resulting from exposure to extreme loadings such as earthquake, vessel/truck impact or fire.
- Create GIS (Geographic information system) maps and Geomatics.

Benefits of 3D Scanning

3D model scanning could benefit the design process if:

- Increase effectiveness working with complex parts and shapes.
- Help with design of products to accommodate someone else's part.

- If CAD models are outdated, a 3D scan will provide an updated version
- Replacement of missing or older parts

Entertainment

3D scanners are used by the entertainment industry to create digital 3D models for both movies and video games. In cases where a real-world equivalent of a model exists, it is much faster to scan the real-world object than to manually create a model using 3D modeling software. Frequently, artists sculpt physical models of what they want and scan them into digital form rather than directly creating digital models on a computer.

Reverse Engineering

Reverse engineering of a mechanical component requires a precise digital model of the objects to be reproduced. Rather than a set of points a precise digital model can be represented by a polygon mesh, a set of flat or curved NURBS surfaces, or ideally for mechanical components, a CAD solid model. A 3D scanner can be used to digitize free-form or gradually changing shaped components as well as prismatic geometries whereas a coordinate measuring machine is usually used only to determine simple dimensions of a highly prismatic model. These data points are then processed to create a usable digital model, usually using specialized reverse engineering software.

Cultural Heritage

There have been many research projects undertaken via the scanning of historical sites and artifacts both for documentation and analysis purposes. The combined use

of 3D scanning and 3D printing technologies allows the replication of real objects without the use of traditional plaster casting techniques, that in many cases can be too invasive for being performed on precious or delicate cultural heritage artifacts. In the side figure the gargoyle model on the left was digitally acquired by using a 3D scanner and the produced 3D data was processed using MeshLab. The resulting digital 3D model, shown in the screen of the laptop, was used by a rapid prototyping machine to create a real resin replica of original object.

Michelangelo

In 1999, two different research groups started scanning Michelangelo's statues. Stanford University with a group led by Marc Levoy used a custom laser triangulation scanner built by Cyberware to scan Michelangelo's statues in Florence, notably the David, the Prigioni and the four statues in The Medici Chapel. The scans produced a data point density of one sample per 0.25 mm, detailed enough to see Michelangelo's chisel marks. These detailed scans produced a huge amount of data (up to 32 gigabytes) and processing the data from his scans took 5 months. Approximately in the same period a research group from IBM, led by H. Rushmeier and F. Bernardini scanned the Pietà of Florence acquiring both geometric and color details. The digital model, result of the Stanford scanning campaign, was thoroughly used in the 2004 subsequent restoration of the statue.

Monticello

In 2002, David Luebke, et al. scanned Thomas Jefferson's Monticello. A commercial time of flight laser scanner, the

DeltaSphere 3000, was used. The scanner data was later combined with color data from digital photographs to create the Virtual Monticello, and the Jefferson's Cabinet exhibits in the New Orleans Museum of Art in 2003. The Virtual Monticello exhibit simulated a window looking into Jefferson's Library. The exhibit consisted of a rear projection display on a wall and a pair of stereo glasses for the viewer. The glasses, combined with polarized projectors, provided a 3D effect. Position tracking hardware on the glasses allowed the display to adapt as the viewer moves around, creating the illusion that the display is actually a hole in the wall looking into Jefferson's Library. The Jefferson's Cabinet exhibit was a barrier stereogram (essentially a non-active hologram that appears different from different angles) of Jefferson's Cabinet

Cuneiform Tablets

In 2003, Subodh Kumar, et al. undertook the 3D scanning of ancient cuneiform tablets. Again, a laser triangulation scanner was used. The tablets were scanned on a regular grid pattern at a resolution of 0.025 mm.

Kasubi Tombs

A 2009 CyArk 3D scanning project at Uganda's historic Kasubi Tombs, a UNESCO World Heritage Site, using a Leica HDS 4500, produced detailed architectural models of Muzibu Azaala Mpanga, the main building at the complex and tomb of the Kabakas (Kings) of Uganda. A fire on March 16, 2010, burned down much of the Muzibu Azaala Mpanga structure, and reconstruction work is likely to lean heavily upon the dataset produced by the 3D scan mission.

"Plastico di Roma antica"

In 2005, Gabriele Guidi, et al. scanned the "Plastico di Roma antica", a model of Rome created in the last century. Neither the triangulation method, nor the time of flight method satisfied the requirements of this project because the item to be scanned was both large and contained small details.

They found though, that a modulated light scanner was able to provide both the ability to scan an object the size of the model and the accuracy that was needed. The modulated light scanner was supplemented by a triangulation scanner which was used to scan some parts of the model.

Medical CAD/CAM

3D scanners are used in order to capture the 3D shape of a patient in orthotics and dentistry. It gradually supplants tedious plaster cast. CAD/CAM software are then used to design and manufacture the orthosis, prosthesis or dental implants. Many Chairside dental CAD/CAM systems and Dental Laboratory CAD/CAM systems use 3D Scanner technologies to capture the 3D surface of a dental preparation (either *in vivo* or *in vitro*), in order to produce a restoration digitally using CAD software and ultimately produce the final restoration using a CAM technology (such as a CNC milling machine, or 3D printer).

The chairside systems are designed to facilitate the 3D scanning of a preparation *in vivo* and produce the restoration (such as a Crown, Onlay, Inlay or Veneer).

Quality Assurance / Industrial Metrology

The digitalization of real-world objects is of vital importance in various application domains. This method is especially applied in industrial quality assurance to measure the geometric dimension accuracy. Industrial processes such as assembly are complex, highly automated and typically based on CAD (Computer Aided Design) data. The problem is that the same degree of automation is also required for quality assurance. It is, for example, a very complex task to assemble a modern car, since it consists of many parts that must fit together at the very end of the production line. The optimal performance of this process is guaranteed by quality assurance systems. Especially the geometry of the metal parts must be checked in order to assure that they have the correct dimensions, fit together and finally work reliably.

Within highly automated processes, the resulting geometric measures are transferred to machines that manufacture the desired objects. Due to mechanical uncertainties and abrasions, the result may differ from its digital nominal. In order to automatically capture and evaluate these deviations, the manufactured part must be digitized as well. For this purpose, 3D scanners are applied to generate point samples from the object's surface which are finally compared against the nominal data .

The process of comparing 3D data against a CAD model is referred to as CAD-Compare, and can be a useful technique for applications such as determining wear patterns on molds and tooling, determining accuracy of final build, analyzing

gap and flush, or analyzing highly complex sculpted surfaces. At present, laser triangulation scanners, structured light and contact scanning are the predominant technologies employed for industrial purposes, with contact scanning remaining the slowest, but overall most accurate option.

2

Optical Character Recognition

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used to convert books and documents into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation, text-to-speech and text mining to it. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. OCR systems require calibration to read a specific font; early versions needed to be programmed with images of each character, and worked on one font at a time. “Intelligent” systems with a high degree of recognition accuracy for most fonts are now

common. Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components.

History

In 1929 Gustav Tauschek obtained a patent on OCR in Germany, followed by Paul W. Handel who obtained a US patent on OCR in USA in 1933 (U.S. Patent 1,915,993). In 1935 Tauschek was also granted a US patent on his method (U.S. Patent 2,026,329). Tauschek's machine was a mechanical device that used templates and a photodetector.

RCA engineers in 1949 worked on the first primitive computer-type OCR to help blind people for the US Veterans Administration, but instead of converting the printed characters to machine language, their device converted it to machine language and then spoke the letters. It proved far too expensive and was not pursued after testing.

In 1950, David H. Shepard, a cryptanalyst at the Armed Forces Security Agency in the United States, addressed the problem of converting printed messages into machine language for computer processing and built a machine to do this, reported in the Washington Daily News on 27 April 1951 and in the New York Times on 26 December 1953 after his U.S. Patent 2,663,758 was issued. Shepard then founded Intelligent Machines Research Corporation (IMR), which went on to deliver the world's first several OCR systems used in commercial operation. The first commercial system was installed at the Reader's Digest in 1955.

The second system was sold to the Standard Oil Company for reading credit card imprints for billing purposes. Other systems sold by IMR during the late 1950s included a bill stub reader to the Ohio Bell Telephone Company and a page scanner to the United States Air Force for reading and transmitting by teletype typewritten messages.

IBM and others were later licensed on Shepard's OCR patents. In about 1965 Reader's Digest and RCA collaborated to build an OCR Document reader designed to digitise the serial numbers on Reader's Digest coupons returned from advertisements. The fonts used on the documents were printed by an RCA Drum printer using the OCR-A font.

The reader was connected directly to an RCA 301 computer (one of the first solid state computers). This reader was followed by a specialised document reader installed at TWA where the reader processed Airline Ticket stock. The readers processed documents at a rate of 1,500 documents per minute, and checked each document, rejecting those it was not able to process correctly. The product became part of the RCA product line as a reader designed to process "Turn around Documents" such as those utility and insurance bills returned with payments. The United States Postal Service has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rabinow.

The first use of OCR in Europe was by the British General Post Office (GPO). In 1965 it began planning an entire banking system, the National Giro, using OCR technology, a process that revolutionized bill payment systems in the UK. Canada Post has been using OCR

systems since 1971. OCR systems read the name and address of the addressee at the first mechanised sorting center, and print a routing bar code on the envelope based on the postal code. To avoid confusion with the human-readable address field which can be located anywhere on the letter, special ink (orange in visible light) is used that is clearly visible under ultraviolet light. Envelopes may then be processed with equipment based on simple barcode readers. In 1974 Ray Kurzweil started the company Kurzweil Computer Products, Inc. and led development of the first omni-font optical character recognition system — a computer programme capable of recognizing text printed in any normal font. He decided that the best application of this technology would be to create a reading machine for the blind, which would allow blind people to have a computer read text to them out loud.

This device required the invention of two enabling technologies — the CCD flatbed scanner and the text-to-speech synthesizer. On January 13, 1976 the successful finished product was unveiled during a widely-reported news conference headed by Kurzweil and the leaders of the National Federation of the Blind. In 1978 Kurzweil Computer Products began selling a commercial version of the optical character recognition computer programme. LexisNexis was one of the first customers, and bought the programme to upload paper legal and news documents onto its nascent online databases. Two years later, Kurzweil sold his company to Xerox, which had an interest in further commercializing paper-to-computer text conversion. Kurzweil Computer Products became a subsidiary of Xerox known as Scansoft,

now Nuance Communications. 1992-1996 Commissioned by the U.S. Department of Energy(DOE), Information Science Research Institute(ISRI) conducted the most authoritative of the Annual Test of OCR Accuracy for 5 consecutive years in the mid-90s. Information Science Research Institute(ISRI) is a research and development unit of University of Nevada, Las Vegas. ISRI was established in 1990 with funding from the U.S. Department of Energy. Its mission is to foster the improvement of automated technologies for understanding machine printed documents.

OCR Software

Desktop & Server OCR Software

OCR Software and ICR Software technology are analytical artificial intelligence systems that consider sequences of characters rather than whole words or phrases. Based on the analysis of sequential lines and curves, OCR and ICR make 'best guesses' at characters using database look-up tables to closely associate or match the strings of characters that form words.

WebOCR & OnlineOCR

With IT technology development, the platform for people to use software has been changed from single PC platform to multi-platforms such as PC +Web-based+ Cloud Computing + Mobile equipments. After 30 years development, OCR software started to adapt to new application requirements. WebOCR also known as OnlineOCR or Web-based OCR service, has been a new trend to meet larger volume and larger group of users after 30 years development of the desktop OCR. Internet and broad band technologies

have made WebOCR & OnlineOCR practically available to both individual users and enterprise customers. Since 2000, some major OCR vendors began offering WebOCR & Online software, a number of new entrants companies to seize the opportunity to develop innovative Web-based OCR service, some of which are free of charge services.

Current State of OCR Technology

Recognition of Latin-script, typewritten text is still not 100% accurate even where clear imaging is available. One study based on recognition of 19th and early 20th century newspaper pages concluded that character-by-character OCR accuracy for commercial OCR software varied from 71% to 98%; total accuracy can only be achieved by human review.

Other areas—including recognition of hand printing, cursive handwriting, and printed text in other scripts (especially those East Asian language characters which have many strokes for a single character)—are still the subject of active research. Accuracy rates can be measured in several ways, and how they are measured can greatly affect the reported accuracy rate. For example, if word context (basically a lexicon of words) is not used to correct software finding non-existent words, a character error rate of 1% (99% accuracy) may result in an error rate of 5% (95% accuracy) or worse if the measurement is based on whether each whole word was recognized with no incorrect letters.

On-line character recognition is sometimes confused with Optical Character Recognition. OCR is an instance of off-line character recognition, where the system recognizes the

fixed *static shape* of the character, while on-line character recognition instead recognizes the *dynamic motion* during handwriting. For example, on-line recognition, such as that used for gestures in the Penpoint OS or the Tablet PC can tell whether a horizontal mark was drawn right-to-left, or left-to-right. On-line character recognition is also referred to by other terms such as dynamic character recognition, real-time character recognition, and Intelligent Character Recognition or ICR.

On-line systems for recognizing hand-printed text on the fly have become well-known as commercial products in recent years. Among these are the input devices for personal digital assistants such as those running Palm OS. The Apple Newton pioneered this product. The algorithms used in these devices take advantage of the fact that the order, speed, and direction of individual lines segments at input are known.

Also, the user can be retrained to use only specific letter shapes. These methods cannot be used in software that scans paper documents, so accurate recognition of hand-printed documents is still largely an open problem. Accuracy rates of 80% to 90% on neat, clean hand-printed characters can be achieved, but that accuracy rate still translates to dozens of errors per page, making the technology useful only in very limited applications.

Recognition of cursive text is an active area of research, with recognition rates even lower than that of hand-printed text. Higher rates of recognition of general cursive script will likely not be possible without the use of contextual or grammatical information. For example, recognizing entire

words from a dictionary is easier than trying to parse individual characters from script. Reading the *Amount* line of a cheque (which is always a written-out number) is an example where using a smaller dictionary can increase recognition rates greatly. Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy. The shapes of individual cursive characters themselves simply do not contain enough information to accurately (greater than 98%) recognise all handwritten cursive script.

It is necessary to understand that OCR technology is a basic technology also used in advanced scanning applications. Due to this, an advanced scanning solution can be unique and patented and not easily copied despite being based on this basic OCR technology.

For more complex recognition problems, intelligent character recognition systems are generally used, as artificial neural networks can be made indifferent to both affine and non-linear transformations.

A technique which is having considerable success in recognising difficult words and character groups within documents generally amenable to computer OCR is to submit them automatically to humans in the reCAPTCHA system.

In 2010, OCR technology was used in a video mode with the release of the iPhone translation application Word Lens that achieved an increase in accuracy by integrating information from multiple frames.

Charge-Coupled Device

A charge-coupled device (CCD) is a device for the movement of electrical charge, usually from within the device to an area where the charge can be manipulated, for example conversion into a digital value. This is achieved by "shifting" the signals between stages within the device one at a time. CCDs move charge between capacitive *bins* in the device, with the shift allowing for the transfer of charge between bins.

Often the device is integrated with an image sensor, such as a photoelectric device to produce the charge that is being read, thus making the CCD a major technology for digital imaging. Although CCDs are not the only technology to allow for light detection, CCDs are widely used in professional, medical, and scientific applications where high-quality image data is required.

History

The charge-coupled device was invented in 1969 at AT&T Bell Labs by Willard Boyle and George E. Smith. The lab was working on semiconductor bubble memory when Boyle and Smith conceived of the design of what they termed, in their notebook, "Charge 'Bubble' Devices". A description of how the device could be used as a shift register and as a linear and area imaging devices was described in this first entry. The essence of the design was the ability to transfer charge along the surface of a semiconductor from one storage capacitor to the next. The concept was similar in principle to the bucket-brigade device (BBD), which was developed at Philips Research Labs during the late 1960's.

The initial paper describing the concept listed possible uses as a memory, a delay line, and an imaging device. The first experimental device demonstrating the principle was a row of closely spaced metal squares on an oxidized silicon surface electrically accessed by wire bonds.

The first working CCD made with integrated circuit technology was a simple 8-bit shift register. This device had input and output circuits and was used to demonstrate its use as a shift register and as a crude eight pixel linear imaging device. Development of the device progressed at a rapid rate. By 1971, Bell researchers Michael F. Tompsett et al. were able to capture images with simple linear devices.

Several companies, including Fairchild Semiconductor, RCA and Texas Instruments, picked up on the invention and began development programmes. Fairchild's effort, led by ex-Bell researcher Gil Amelio, was the first with commercial devices, and by 1974 had a linear 500-element device and a 2-D 100 x 100 pixel device. The first KH-11 KENNAN reconnaissance satellite equipped with charge-coupled device array technology for imaging was launched in December 1976. Under the leadership of Kazuo Iwama, Sony also started a big development effort on CCDs involving a significant investment. Eventually, Sony managed to mass produce CCDs for their camcorders. Before this happened, Iwama died in August 1982; subsequently, a CCD chip was placed on his tombstone to acknowledge his contribution.

In January 2006, Boyle and Smith were awarded the National Academy of Engineering Charles Stark Draper Prize, and in 2009 they were awarded the Nobel Prize for Physics, for their work on the CCD.

Basics of Operation

In a CCD for capturing images, there is a photoactive region (an epitaxial layer of silicon), and a transmission region made out of a shift register (the CCD, properly speaking).

An image is projected through a lens onto the capacitor array (the photoactive region), causing each capacitor to accumulate an electric charge proportional to the light intensity at that location. A one-dimensional array, used in line-scan cameras, captures a single slice of the image, while a two-dimensional array, used in video and still cameras, captures a two-dimensional picture corresponding to the scene projected onto the focal plane of the sensor.

Once the array has been exposed to the image, a control circuit causes each capacitor to transfer its contents to its neighbor (operating as a shift register). The last capacitor in the array dumps its charge into a charge amplifier, which converts the charge into a voltage.

By repeating this process, the controlling circuit converts the entire contents of the array in the semiconductor to a sequence of voltages.

In a digital device, these voltages are then sampled, digitized, and usually stored in memory; in an analog device (such as an analog video camera), they are processed into a continuous analog signal (e.g. by feeding the output of the charge amplifier into a low-pass filter) which is then processed and fed out to other circuits for transmission, recording, or other processing.

Detailed Physics of Operation

The photoactive region of the CCD is, generally, an epitaxial layer of silicon. It has a doping of p+ (Boron) and is grown upon a substrate material, often p++. In buried channel devices, the type of design utilized in most modern CCDs, certain areas of the surface of the silicon are ion implanted with phosphorus, giving them an n-doped designation. This region defines the channel in which the photogenerated charge packets will travel. The gate oxide, i.e. the capacitor dielectric, is grown on top of the epitaxial layer and substrate. Later on in the process polysilicon gates are deposited by chemical vapor deposition, patterned with photolithography, and etched in such a way that the separately phased gates lie perpendicular to the channels. The channels are further defined by utilization of the LOCOS process to produce the channel stop region. Channel stops are thermally grown oxides that serve to isolate the charge packets in one column from those in another.

These channel stops are produced before the polysilicon gates are, as the LOCOS process utilizes a high temperature step that would destroy the gate material. The channels stops are parallel to, and exclusive of, the channel, or "charge carrying", regions. Channel stops often have a p+ doped region underlying them, providing a further barrier to the electrons in the charge packets (this discussion of the physics of CCD devices assumes an electron transfer device, though hole transfer is possible). The clocking of the gates, alternately high and low, will forward and reverse bias to the diode that is provided by the buried channel (n-doped) and the epitaxial layer (p-doped).

This will cause the CCD to deplete, near the p-n junction and will collect and move the charge packets beneath the gates—and within the channels—of the device. CCD manufacturing and operation can be optimized for different uses. The above process describes a frame transfer CCD. While CCDs may be manufactured on a heavily doped p++ wafer it is also possible to manufacture a device inside p-wells that have been placed on an n-wafer.

This second method, reportedly, reduces smear, dark current, and infrared and red response. This method of manufacture is used in the construction of interline transfer devices. Another version of CCD is called a peristaltic CCD. In a peristaltic charge-coupled device, the charge packet transfer operation is analogous to the peristaltic contraction and dilation of the digestive system. The peristaltic CCD has an additional implant that keeps the charge away from the silicon/silicon dioxide interface and generates a large lateral electric field from one gate to the next. This provides an additional driving force to aid in transfer of the charge packets.

Architecture

The CCD image sensors can be implemented in several different architectures. The most common are full-frame, frame-transfer, and interline. The distinguishing characteristic of each of these architectures is their approach to the problem of shuttering. In a full-frame device, all of the image area is active, and there is no electronic shutter.

A mechanical shutter must be added to this type of sensor or the image smears as the device is clocked or read

out. With a frame-transfer CCD, half of the silicon area is covered by an opaque mask (typically aluminum). The image can be quickly transferred from the image area to the opaque area or storage region with acceptable smear of a few percent. That image can then be read out slowly from the storage region while a new image is integrating or exposing in the active area. Frame-transfer devices typically do not require a mechanical shutter and were a common architecture for early solid-state broadcast cameras. The downside to the frame-transfer architecture is that it requires twice the silicon real estate of an equivalent full-frame device; hence, it costs roughly twice as much. The interline architecture extends this concept one step further and masks every other column of the image sensor for storage.

In this device, only one pixel shift has to occur to transfer from image area to storage area; thus, shutter times can be less than a microsecond and smear is essentially eliminated. The advantage is not free, however, as the imaging area is now covered by opaque strips dropping the fill factor to approximately 50 percent and the effective quantum efficiency by an equivalent amount. Modern designs have addressed this deleterious characteristic by adding microlenses on the surface of the device to direct light away from the opaque regions and on the active area. Microlenses can bring the fill factor back up to 90 percent or more depending on pixel size and the overall system's optical design. The choice of architecture comes down to one of utility. If the application cannot tolerate an expensive, failure-prone, power-intensive mechanical shutter, an interline device is the right choice. Consumer snap-shot cameras

have used interline devices. On the other hand, for those applications that require the best possible light collection and issues of money, power and time are less important, the full-frame device is the right choice. Astronomers tend to prefer full-frame devices. The frame-transfer falls in between and was a common choice before the fill-factor issue of interline devices was addressed. Today, frame-transfer is usually chosen when an interline architecture is not available, such as in a back-illuminated device. CCDs containing grids of pixels are used in digital cameras, optical scanners, and video cameras as light-sensing devices.

They commonly respond to 70 percent of the incident light (meaning a quantum efficiency of about 70 percent) making them far more efficient than photographic film, which captures only about 2 percent of the incident light. Most common types of CCDs are sensitive to near-infrared light, which allows infrared photography, night-vision devices, and zero lux (or near zero lux) video-recording/ photography.

For normal silicon-based detectors, the sensitivity is limited to 1.1 ìm. One other consequence of their sensitivity to infrared is that infrared from remote controls often appears on CCD-based digital cameras or camcorders if they do not have infrared blockers. Cooling reduces the array's dark current, improving the sensitivity of the CCD to low light intensities, even for ultraviolet and visible wavelengths. Professional observatories often cool their detectors with liquid nitrogen to reduce the dark current, and therefore the thermal noise, to negligible levels.

Use in Astronomy

Due to the high quantum efficiencies of CCDs, linearity of their outputs (one count for one photon of light), ease of use compared to photographic plates, and a variety of other reasons, CCDs were very rapidly adopted by astronomers for nearly all UV-to-infrared applications. Thermal noise and cosmic rays may alter the pixels in the CCD array.

To counter such effects, astronomers take several exposures with the CCD shutter closed and opened. The average of images taken with the shutter closed is necessary to lower the random noise. Once developed, the *dark frame* average image is then subtracted from the open-shutter image to remove the dark current and other systematic defects (dead pixels, hot pixels, etc.) in the CCD. The Hubble Space Telescope, in particular, has a highly developed series of steps ("data reduction pipeline") to convert the raw CCD data to useful images. See the references for a more in-depth description of the steps in astronomical CCD image-data correction and processing. CCD cameras used in astrophotography often require sturdy mounts to cope with vibrations from wind and other sources, along with the tremendous weight of most imaging platforms. To take long exposures of galaxies and nebulae, many astronomers use a technique known as auto-guiding. Most autoguiders use a second CCD chip to monitor deviations during imaging. This chip can rapidly detect errors in tracking and command the mount motors to correct for them. An interesting unusual astronomical application of CCDs, called *drift-scanning*, uses a CCD to make a fixed telescope behave like a tracking

telescope and follow the motion of the sky. The charges in the CCD are transferred and read in a direction parallel to the motion of the sky, and at the same speed. In this way, the telescope can image a larger region of the sky than its normal field of view. The Sloan Digital Sky Survey is the most famous example of this, using the technique to produce the largest uniform survey of the sky yet accomplished. In addition to astronomy, CCDs are also used in laboratory analytical instrumentation such as monochromators, spectrometers, and N-slit laser interferometers.

Color Cameras

Digital color cameras generally use a Bayer mask over the CCD. Each square of four pixels has one filtered red, one blue, and two green (the human eye is more sensitive to green than either red or blue). The result of this is that luminance information is collected at every pixel, but the color resolution is lower than the luminance resolution.

Better color separation can be reached by three-CCD devices (3CCD) and a dichroic beam splitter prism, that splits the image into red, green and blue components. Each of the three CCDs is arranged to respond to a particular color. Most professional video camcorders, and some semi-professional camcorders, use this technique. Another advantage of 3CCD over a Bayer mask device is higher quantum efficiency (and therefore higher light sensitivity for a given aperture size). This is because in a 3CCD device most of the light entering the aperture is captured by a sensor, while a Bayer mask absorbs a high proportion (about 2/3) of the light falling on each CCD pixel.

For still scenes, for instance in microscopy, the resolution of a Bayer mask device can be enhanced by Microscanning technology. During the process of color co-site sampling, several frames of the scene are produced. Between acquisitions, the sensor is moved in pixel dimensions, so that each point in the visual field is acquired consecutively by elements of the mask that are sensitive to the red, green and blue components of its color. Eventually every pixel in the image has been scanned at least once in each color and the resolution of the three channels become equivalent (the resolutions of red and blue channels are quadrupled while the green channel is doubled).

Sensor Sizes

Sensors (CCD / CMOS) are often referred to with an inch fraction designation such as 1/1.8" or 2/3" called the optical format. This measurement actually originates back in the 1950s and the time of Vidicon tubes. Compact digital cameras and Digicams typically have much smaller sensors than a digital SLR and are thus less sensitive to light and inherently more prone to noise.

Frame Transfer CCD

A frame transfer CCD is a specialized CCD, often used in astronomy and some professional video cameras, designed for high exposure efficiency and correctness.

The normal functioning of a CCD, astronomical or otherwise, can be divided into two phases: exposure and readout. During the first phase, the CCD passively collects incoming photons, storing electrons in its cells. After the exposure time is passed, the cells are read out one line at

a time. During the readout phase, cells are shifted down the entire area of the CCD. While they are shifted, they continue to collect light. Thus, if the shifting is not fast enough, errors can result from light that falls on a cell holding charge during the transfer. These errors are referred to as "vertical smear" and cause a strong light source to create a vertical line above and below its exact location. In addition, the CCD cannot be used to collect light while it is being read out. Unfortunately, a faster shifting requires a faster readout, and a faster readout can introduce errors in the cell charge measurement, leading to a higher noise level. A frame transfer CCD solves both problems: it has a hidden, not normally used, area containing as many cells as the area exposed to light. Typically, this area is covered by a reflective material such as aluminium. When the exposure time is up, the cells are transferred very rapidly to the hidden area. Here, safe from any incoming light, cells can be read out at any speed one deems necessary to correctly measure the cells' charge. At the same time, the exposed part of the CCD is collecting light again, so no delay occurs between successive exposures.

The disadvantage of such a CCD is the higher cost: the cell area is basically doubled, and more complex control electronics are needed.

Intensified Charge-coupled Device

An intensified charge-coupled device (ICCD) is a CCD that is optically connected to an image intensifier that is mounted in front of the CCD. An image intensifier includes three functional elements: a photocathode, a micro-channel

plate (MCP) and a phosphor screen. These three elements are mounted one close behind the other in the mentioned sequence.

The photons which are coming from the light source fall onto the photocathode, thereby generating photoelectrons. The photoelectrons are accelerated towards the MCP by an electrical control voltage, applied between photocathode and MCP. The electrons are multiplied inside of the MCP and thereafter accelerated towards the phosphor screen. The phosphor screen finally converts the multiplied electrons back to photons which are guided to the CCD by a fiber optic or a lens.

An image intensifier inherently includes a shutter functionality: If the control voltage between the photocathode and the MCP is reversed, the emitted photoelectrons are not accelerated towards the MCP but return to the photocathode. Thus, no electrons are multiplied and emitted by the MCP, no electrons are going to the phosphor screen and no light is emitted from the image intensifier. In this case no light falls onto the CCD, which means that the shutter is closed. The process of reversing the control voltage at the photocathode is called gating and therefore ICCDs are also called gateable CCD cameras.

Beside of the extremely high sensitivity of ICCD cameras, which enable single photon detection, the gateability is one of the major advantages of the ICCD over the EMCCD cameras. The highest performing ICCD cameras enable shutter times as short as 200 picoseconds. ICCD cameras are in general somewhat higher in price than EMCCD

cameras because they need the expensive image intensifier. On the other hand EMCCD cameras need a cooling system to cool the EMCCD chip down to temperatures around 170 K. This cooling system adds additional costs to the EMCCD camera and often yields heavy condensation problems in the application. ICCDs are used in night vision devices and in a large variety of scientific applications.

3

Digital Camera

A digital camera (or digicam) is a camera that takes video or still photographs, or both, digitally by recording images via an electronic image sensor. Most 21st century cameras are digital. Digital cameras can do things film cameras cannot: displaying images on a screen immediately after they are recorded, storing thousands of images on a single small memory device, and deleting images to free storage space. The majority, including most compact cameras, can record moving video with sound as well as still photographs. Some can crop and stitch pictures and perform other elementary image editing. Some have a GPS receiver built in, and can produce Geotagged photographs. The optical system works the same as in film cameras, typically using a lens with a variable diaphragm to focus light onto an image pickup device. The diaphragm and shutter admit the correct amount of light to the imager, just as with film but

the image pickup device is electronic rather than chemical. Most digicams, apart from camera phones and a few specialized types, have a standard tripod screw. Digital cameras are incorporated into many devices ranging from PDAs and mobile phones (called camera phones) to vehicles. The Hubble Space Telescope and other astronomical devices are essentially specialized digital cameras.

Types of Digital Cameras

Digital cameras are made in a wide range of sizes, prices and capabilities. The majority are camera phones, operated as a mobile application through the cellphone menu. Professional photographers and many amateurs use larger, more expensive digital single-lens reflex cameras (DSLR) for their greater versatility. Between these extremes lie digital compact cameras and bridge digital cameras that "bridge" the gap between amateur and professional cameras. Specialized cameras including multispectral imaging equipment and astrographs continue to serve the scientific, military, medical and other special purposes for which digital photography was invented.

Compact Digital Cameras

Compact cameras are designed to be tiny and portable and are particularly suitable for casual and "snapshot" use, thus are also called point-and-shoot cameras. The smallest, generally less than 20 mm thick, are described as *subcompacts* or "ultra-compacts" and some are nearly credit card size. Most, apart from ruggedized or water-resistant models, incorporate a retractable lens assembly allowing a thin camera to have a moderately long focal length and thus

fully exploit an image sensor larger than that on a camera phone, and a mechanized lens cap to cover the lens when retracted. The retracted and capped lens is protected from keys, coins and other hard objects, thus making a thin, pocketable package. Subcompacts commonly have one lug and a short wrist strap which aids extraction from a pocket, while thicker compacts may have two lugs for attaching a neck strap.

Compact cameras are usually designed to be easy to use, sacrificing advanced features and picture quality for compactness and simplicity; images can usually only be stored using lossy compression (JPEG). Most have a built-in flash usually of low power, sufficient for nearby subjects. Live preview is almost always used to frame the photo. Most have limited motion picture capability. Compacts often have macro capability and zoom lenses but the zoom range is usually less than for bridge and DSLR cameras. Generally a contrast-detect autofocus system, using the image data from the live preview feed of the main imager, focuses the lens. Typically, these cameras incorporate a nearly-silent leaf shutter into their lenses. For lower cost and smaller size, these cameras typically use image sensors with a diagonal of approximately 6 mm, corresponding to a crop factor around 6. This gives them weaker low-light performance, greater depth of field, generally closer focusing ability, and smaller components than cameras using larger sensors. Some recent compact digital cameras can now able to take 3D still photos. These new 3D compact digital cameras can capture 3D panoramic photos for play back on a 3D TV.

Bridge Cameras

Bridge are higher-end digital cameras that physically and ergonomically resemble DSLRs and share with them some advanced features, but share with compacts the use of a fixed lens and a small sensor. Like compacts, most use live preview to frame the image. Their autofocus uses the same contrast-detect mechanism, but many bridge cameras have a manual focus mode, in some cases using a separate focus ring, for greater control.

Due to the combination of big physical size but a small sensor, many of these cameras have very highly specified lenses with large zoom range and fast aperture, partially compensating for the inability to change lenses.

To compensate for the lesser sensitivity of their small sensors, these cameras almost always include an image stabilization system to enable longer handheld exposures. The highest zoom lens so far on a bridge camera is on the Nikon Coolpix P500 digital camera, which encompasses an equivalent of a super wide to ultra-telephoto 22.5-810 mm (36x).

These cameras are sometimes marketed as and confused with digital SLR cameras since the appearance is similar. Bridge cameras lack the reflex viewing system of DSLRs, are usually fitted with fixed (non-interchangeable) lenses (although some have a lens thread to attach accessory wide-angle or telephoto converters), and can usually take movies with sound. The scene is composed by viewing either the liquid crystal display or the electronic viewfinder (EVF). Most have a longer shutter lag than a true dSLR, but they

are capable of good image quality (with sufficient light) while being more compact and lighter than DSLRs. High-end models of this type have comparable resolutions to low and mid-range DSLRs. Many of these cameras can store images in a Raw image format, or processed and JPEG compressed, or both. The majority have a built-in flash similar to those found in DSLRs.

In bright sun, the quality difference between a good compact camera and a digital SLR is minimal but bridgecams are more portable, cost less and have a similar zoom ability to dSLR. Thus a Bridge camera may better suit outdoor daytime activities, except when seeking professional-quality photos. In low light conditions and/or at ISO equivalents above 800, most bridge cameras (or megazooms) lack in image quality when compared to even entry level DSLRs.

Mirrorless Interchangeable Lens Camera

In late 2008 a new type of camera emerged, combining the larger sensors and interchangeable lenses of DSLRs with the live preview viewing system of compact cameras, either through an electronic viewfinder or on the rear LCD. These are simpler and more compact than DSLRs due to the removal of the mirror box, and typically emulate the handling and ergonomics of either DSLRs or compacts.

The system is use by Micro Four Thirds, borrowing components from the Four Thirds DSLR system. The Ricoh GXR of 2009 puts the sensor and other electronic components in the interchangeable sensor lens unit rather than in the camera body. The first interchangeable 3D lens Lumix G 12.5mm/F12 (H-FT012) has been announced by

Panasonic. It use two lenses quite close together in one lens module adaptor and record both 3D and 2D pictures altogether. The lens module is compatible with Panasonic Lumix DMC-GH2.

Digital Single Lens Reflex Cameras

Digital single-lens reflex cameras (DSLRs) are digital cameras based on film single-lens reflex cameras (SLRs). They take their name from their unique viewing system, in which a mirror reflects light from the lens through a separate optical viewfinder. In order to capture an image the mirror is flipped out of the way, allowing light to fall on the imager. Since no light reaches the imager during framing, autofocus is accomplished using specialized sensors in the mirror box itself. Most 21st century DSLRs also have a "live view" mode that emulates the live preview system of compact cameras, when selected. These cameras have much larger sensors than the other types, typically 18 mm to 36 mm on the diagonal (crop factor 2, 1.6, or 1). This gives them superior low-light performance, less depth of field at a given aperture, and a larger size.

They make use of interchangeable lenses; each major DSLR manufacturer also sells a line of lenses specifically intended to be used on their cameras. This allows the user to select a lens designed for the application at hand: wide-angle, telephoto, low-light, etc. So each lens does not require its own shutter, DSLRs use a focal-plane shutter in front of the imager, behind the mirror. The mirror flipping out of the way at the moment of exposure makes a distinctive "clack" sound.

Digital Rangefinders

A rangefinder is a user-operated optical mechanism to measure subject distance once widely used on film cameras. Most digital cameras measure subject distance automatically using electro-optical techniques, but it is not customary to say that they have a rangefinder.

Line-scan Camera Systems

A line-scan camera is a camera device containing a line-scan image sensor chip, and a focusing mechanism. These cameras are almost solely used in industrial settings to capture an image of a constant stream of moving material. Unlike video cameras, line-scan cameras use a single array of pixel sensors, instead of a matrix of them.

Data coming from the line-scan camera has a frequency, where the camera scans a line, waits, and repeats. The data coming from the line-scan camera is commonly processed by a computer, to collect the one-dimensional line data and to create a two-dimensional image. The collected two-dimensional image data is then processed by image-processing methods for industrial purposes.

Line-scan technology is capable of capturing data extremely fast, and at very high image resolutions. Usually under these conditions, resulting collected image data can quickly exceed 100 MB in a fraction of a second. Line-scan-camera–based integrated systems, therefore are usually designed to streamline the camera's output in order to meet the system's objective, using computer technology which is also affordable.

Line-scan cameras intended for the parcel handling industry can integrate adaptive focusing mechanisms to scan six sides of any rectangular parcel in focus, regardless of angle, and size. The resulting 2-D captured images could contain, but are not limited to 1D and 2D barcodes, address information, and any pattern that can be processed via image processing methods. Since the images are 2-D, they are also human-readable and can be viewable on a computer screen. Advanced integrated systems include video coding, optical character recognition (OCR) and finish-line cameras for high speed sports.

Integration

Many devices include digital cameras built into or integrated into them. For example, mobile phones often include digital cameras; those that do are known as camera phones. Other small electronic devices (especially those used for communication) such as PDAs, laptops and BlackBerry devices often contain an integral digital camera, and most 21st century camcorders can also make still pictures. Due to the limited storage capacity and general emphasis on convenience rather than image quality, the vast majority of these integrated or converged devices store images in the lossy but compact JPEG file format. Mobile phones incorporating digital cameras were introduced in Japan in 2001 by J-Phone. In 2003 camera phones outsold stand-alone digital cameras, and in 2006 they outsold all film-based cameras and digital cameras combined. These camera phones reached a billion devices sold in only five years, and by 2007 more than half of the installed base of

all mobile phones were camera phones. Sales of separate cameras peaked in 2008.

Integrated cameras tend to be at the very lowest end of the scale of digital cameras in technical specifications, such as resolution, optical quality, and ability to use accessories. With rapid development, however, the gap between mainstream compact digital cameras and camera phones is closing, and high-end camera phones are competitive with low-end stand-alone digital cameras of the same generation.

Conversion of Film Cameras to Digital

When digital cameras became common, a question many photographers asked was whether their film cameras could be converted to digital. The answer was yes and no. For the majority of 35 mm film cameras the answer is no, the reworking and cost would be too great, especially as lenses have been evolving as well as cameras. For most a conversion to digital, to give enough space for the electronics and allow a liquid crystal display to preview, would require removing the back of the camera and replacing it with a custom built digital unit.

Many early professional SLR cameras, such as the Kodak DCS series, were developed from 35 mm film cameras. The technology of the time, however, meant that rather than being digital "backs" the bodies of these cameras were mounted on large, bulky digital units, often bigger than the camera portion itself. These were factory built cameras, however, not aftermarket conversions.

A notable exception is the Nikon E2, followed by Nikon E3, using additional optics to convert the 35mm format to a 2/3 CCD-sensor. A few 35 mm cameras have had digital camera backs made by their manufacturer, Leica being a notable example. Medium format and large format cameras (those using film stock greater than 35 mm), have a low unit production, and typical digital backs for them cost over $10,000. These cameras also tend to be highly modular, with handgrips, film backs, winders, and lenses available separately to fit various needs.

The very large sensor these backs use leads to enormous image sizes. For example Phase One's P45 39 MP image back creates a single TIFF image of size up to 224.6 MB, and even greater pixel counts are available. Medium format digitals such as this are geared more towards studio and portrait photography than their smaller DSLR counterparts; the ISO speed in particular tends to have a maximum of 400, versus 6400 for some DSLR cameras. (Canon EOS-1D Mark IV and Nikon D3S have ISO 12800 plus Hi-3 ISO 102400)

History

Digital cameras were developed in the last quarter of the 20th century, from predecessors including video camera tubes

Image Sensors

Image Resolution

The resolution of a digital camera is often limited by the image sensor (typically a CCD or CMOS sensor chip) that

turns light into discrete signals, replacing the job of film in traditional photography. The sensor is made up of millions of "buckets" that essentially count the number of photons that strike the sensor. This means that the brighter the image at a given point on the sensor, the larger the value that is read for that pixel. Depending on the physical structure of the sensor, a color filter array may be used which requires a demosaicing/interpolation algorithm. The number of resulting pixels in the image determines its "pixel count". For example, a 640x480 image would have 307,200 pixels, or approximately 307 kilopixels; a 3872x2592 image would have 10,036,224 pixels, or approximately 10 megapixels.

The pixel count alone is commonly presumed to indicate the resolution of a camera, but this simple figure of merit is a misconception. Other factors impact a sensor's resolution, including sensor size, lens quality, and the organization of the pixels (for example, a monochrome camera without a Bayer filter mosaic has a higher resolution than a typical color camera). Many digital compact cameras are criticized for having excessive pixels.

Sensors can be so small that their 'buckets' can easily overfill; again, resolution of a sensor can become greater than the camera lens could possibly deliver. As the technology has improved, costs have decreased dramatically. Counting the "pixels per dollar" as a basic measure of value for a digital camera, there has been a continuous and steady increase in the number of pixels each dollar buys in a new camera, in accord with the principles of Moore's Law. This predictability of camera prices was first presented in 1998

at the Australian PMA DIMA conference by Barry Hendy and since referred to as "Hendy's Law". Since only a few aspect ratios are commonly used (mainly 4:3 and 3:2), the number of sensor sizes that are useful is limited. Furthermore, sensor manufacturers do not produce every possible sensor size, but take incremental steps in sizes. For example, in 2007 the three largest sensors (in terms of pixel count) used by Canon were the 21.1, 17.9, and 16.6 megapixel CMOS sensors.

Methods of Image Capture

Since the first digital backs were introduced, there have been three main methods of capturing the image, each based on the hardware configuration of the sensor and color filters. The first method is often called *single-shot*, in reference to the number of times the camera's sensor is exposed to the light passing through the camera lens. Single-shot capture systems use either one CCD with a Bayer filter mosaic, or three separate image sensors (one each for the primary additive colors red, green, and blue) which are exposed to the same image via a beam splitter.

The second method is referred to as *multi-shot* because the sensor is exposed to the image in a sequence of three or more openings of the lens aperture. There are several methods of application of the multi-shot technique. The most common originally was to use a single image sensor with three filters (once again red, green and blue) passed in front of the sensor in sequence to obtain the additive color information. Another multiple shot method is called Microscanning. This technique utilizes a single CCD with

a Bayer filter but actually moved the physical location of the sensor chip on the focus plane of the lens to "stitch" together a higher resolution image than the CCD would allow otherwise. A third version combined the two methods without a Bayer filter on the chip.

The third method is called *scanning* because the sensor moves across the focal plane much like the sensor of a desktop scanner. Their *linear* or *tri-linear* sensors utilize only a single line of photosensors, or three lines for the three colors. In some cases, scanning is accomplished by moving the sensor e.g. when using Color co-site sampling or rotate the whole camera; a digital rotating line camera offers images of very high total resolution.

The choice of method for a given capture is determined largely by the subject matter. It is usually inappropriate to attempt to capture a subject that moves with anything but a single-shot system. However, the higher color fidelity and larger file sizes and resolutions available with multi-shot and scanning backs make them attractive for commercial photographers working with stationary subjects and large-format photographs. Dramatic improvements in single-shot cameras and raw image file processing at the beginning of the 21st century made single shot, CCD-based cameras almost completely dominant, even in high-end commercial photography. CMOS-based single shot cameras remained somewhat common.

Filter Mosaics, Interpolation, and Aliasing

Most current consumer digital cameras use a Bayer filter mosaic in combination with an optical anti-aliasing filter to

reduce the aliasing due to the reduced sampling of the different primary-color images. A demosaicing algorithm is used to interpolate color information to create a full array of RGB image data. Cameras that use a beam-splitter single-shot 3CCD approach, three-filter multi-shot approach, Color co-site sampling or Foveon X3 sensor do not use anti-aliasing filters, nor demosaicing.

Firmware in the camera, or a software in a raw converter programme such as Adobe Camera Raw, interprets the raw data from the sensor to obtain a full color image, because the RGB color model requires three intensity values for each pixel: one each for the red, green, and blue (other color models, when used, also require three or more values per pixel). A single sensor element cannot simultaneously record these three intensities, and so a color filter array (CFA) must be used to selectively filter a particular color for each pixel. The Bayer filter pattern is a repeating 2×2 mosaic pattern of light filters, with green ones at opposite corners and red and blue in the other two positions. The high proportion of green takes advantage of properties of the human visual system, which determines brightness mostly from green and is far more sensitive to brightness than to hue or saturation. Sometimes a 4-color filter pattern is used, often involving two different hues of green. This provides potentially more accurate color, but requires a slightly more complicated interpolation process.

The color intensity values not captured for each pixel can be interpolated (or guessed) from the values of adjacent pixels which represent the color being calculated.

Connectivity

Transferring Photos

Many digital cameras can connect directly to a computer to transfer data:

- Early cameras used the PC serial port. USB is now the most widely used method (most cameras are viewable as USB mass storage), though some have a FireWire port. Some cameras use USB PTP mode for connection instead of USB MSC; some offer both modes.
- Other cameras use wireless connections, via Bluetooth or IEEE 802.11 WiFi, such as the Kodak EasyShare One.
- Cameraphones and some high-end stand-alone digital cameras also use cellular networks to connect for sharing images. The most common standard on cellular networks is the MMS Multimedia Messaging Service, commonly called "picture messaging". The second method with smartphones is to send a picture as an email attachment. Many cameraphones do not support email, so this is less common.

A common alternative is the use of a card reader which may be capable of reading several types of storage media, as well as high speed transfer of data to the computer. Use of a card reader also avoids draining the camera battery during the download process, as the device takes power from the USB port. An external card reader allows convenient direct access to the images on a collection of storage media. But if only one storage card is in use, moving it back and

forth between the camera and the reader can be inconvenient. Many computers have a card reader built in, at least for SD cards.

Printing Photos

Many modern cameras support the PictBridge standard, which allows them to send data directly to a PictBridge-capable computer printer without the need for a computer. Wireless connectivity can also provide for printing photos without a cable connection. Polaroid has introduced a printer integrated into its digital camera which creates a small, printed copy of a photo. This is reminiscent of the original instant camera, popularized by Polaroid in 1975.

Displaying Photos

Many digital cameras include a video output port. Usually sVideo, it sends a standard-definition video signal to a television, allowing the user to show one picture at a time. Buttons or menus on the camera allow the user to select the photo, advance from one to another, or automatically send a "slide show" to the TV. HDMI has been adopted by many high-end digital camera makers, to show photos in their high-resolution quality on an HDTV.

In January 2008, Silicon Image announced a new technology for sending video from mobile devices to a television in digital form. MHL sends pictures as a video stream, up to 1080p resolution, and is compatible with HDMI. Some DVD recorders and television sets can read memory cards used in cameras; alternatively several types of flash card readers have TV output capability.

Modes

Many digital cameras have preset modes for different applications. Within the constraints of correct exposure various parameters can be changed, including exposure, aperture, focusing, light metering, white balance, and equivalent sensitivity. For example a portrait might use a wider aperture to render the background out of focus, and would seek out and focus on a human face rather than other image content.

Image Data Storage

Many camera phones and most separate digital cameras use memory cards having flash memory to store image data. The majority of cards for separate cameras are SD format; many are CompactFlash or other formats. Digital cameras have computers inside, hence have internal memory, and many cameras can use some of this internal memory for a limited capacity for pictures that can be transferred to or from the card or through the camera's connections. A few cameras use some other form of removable storage such as Microdrives (very small hard disk drives), CD single (185 MB), and 3.5" floppy disks. Other unusual formats include:

- Onboard flash memory — Cheap cameras and cameras secondary to the device's main use (such as a camera phone)
- PC Card hard drives — early professional cameras (discontinued)
- Thermal printer — known only in one model of camera that printed images immediately rather than storing

Most manufacturers of digital cameras do not provide drivers and software to allow their cameras to work with Linux or other free software. Still, many cameras use the standard USB storage protocol, and are thus easily usable. Other cameras are supported by the gPhoto project.

File Formats

The Joint Photography Experts Group standard (JPEG) is the most common file format for storing image data. Other file types include Tagged Image File Format (TIFF) and various Raw image formats. Many cameras, especially professional or DSLR cameras, support a Raw image format. A raw image is the unprocessed set of pixel data directly from the camera's sensor. They are often saved in formats proprietary to each manufacturer, such as NEF for Nikon, CRW or CR2 for Canon, and MRW for Minolta. Adobe Systems has released the DNG format, a royalty free raw image format which has been adopted by at least 10 camera manufacturers.

Raw files initially had to be processed in specialized image editing programmes, but over time many mainstream editing programmes, such as Google's Picasa, have added support for raw images. Editing raw format images allows more flexibility in settings such as white balance, exposure compensation, color temperature, and so on. In essence raw format allows the photographer to make major adjustments without losing image quality that would otherwise require retaking the picture.

Formats for movies are AVI, DV, MPEG, MOV (often containing motion JPEG), WMV, and ASF (basically the

same as WMV). Recent formats include MP4, which is based on the QuickTime format and uses newer compression algorithms to allow longer recording times in the same space.

Other formats that are used in cameras but not for pictures are the Design Rule for Camera Format (DCF), an ISO specification for the camera's internal file structure and naming, and Digital Print Order Format (DPOF), which dictates what order images are to be printed in and how many copies.

Most cameras include Exif data that provides metadata about the picture. Exif data may include aperture, exposure time, focal length, date and time taken, and location.

Batteries

Digital cameras have high power requirements, and over time have become smaller, resulting in an ongoing need to develop a battery small enough to fit in the camera and yet able to power it for a reasonable length of time.

Two broad types of batteries are in use for digital cameras.

Off-the-shelf

The first type of battery for digital cameras conform to an established off-the-shelf form factor, most commonly AA, CR2, or CR-V3 batteries, with AAA batteries in a handful of cameras. The CR2 and CR-V3 batteries are lithium based, and intended for single use. They are also commonly seen in camcorders. AA batteries are the most common; however, the non-rechargeable alkaline batteries supplied with low-end cameras are capable of providing enough power for only

a very short time in most cameras. They may serve satisfactorily in cameras that are only occasionally used.

Consumers with more than an occasional need use AA Nickel metal hydride batteries (NiMH) instead, which provide an adequate amount of power and are rechargeable. NIMH batteries do not provide as much power per volume as lithium ion batteries, and they also tend to discharge when not used.

To get same power, NiMH Rechargeable battery takes up to two times in volume compare to Li-on Rechargeable Battery, by weight NiMH Rechargeable Battery is three to five times heavier, but by price NiMH Rechargeable Battery is only a half compare to Li-on Rechargeable Battery. Please see Wikipedia: Table of rechargeable battery technologies in Rechargeable battery.

They are available in various ampere-hour (Ah) or milli-ampere-hour (mAh) ratings, which affects how long they last in use. Typically mid-range consumer models and some low end cameras use off-the-shelf batteries; only a very few DSLR cameras accept them (for example, Sigma SD10). Rechargeable RCR-V3 lithium-ion batteries are also available as an alternative to non-rechargeable CR-V3 batteries.

Proprietary

The second type of battery for digital cameras is proprietary battery formats. These are built to a manufacturer's custom specifications, and can be either aftermarket replacement parts or OEM. Almost all proprietary batteries are lithium ion. While they only accept a certain number of recharges before the battery life begins degrading

(typically up to 500 cycles), they provide considerable performance for their size. A result is that at the two ends of the spectrum both high end professional cameras and low end consumer models tend to use lithium ion batteries.

4

Photomultiplier

Photomultiplier tubes (photomultipliers or PMTs for short), members of the class of vacuum tubes, and more specifically phototubes, are extremely sensitive detectors of light in the ultraviolet, visible, and near-infrared ranges of the electromagnetic spectrum. These detectors multiply the current produced by incident light by as much as 100 million times (i.e., 160 dB), in multiple dynode stages, enabling (for example) individual photons to be detected when the incident flux of light is very low. The combination of high gain, low noise, high frequency response, and large area of collection has earned photomultipliers an essential place in nuclear and particle physics, astronomy, medical diagnostics including blood tests, medical imaging, motion picture film scanning (telecine), and high-end image scanners known as drum scanners. Semiconductor devices, particularly avalanche photodiodes, are alternatives to

photomultipliers; however, photomultipliers are uniquely well-suited for applications requiring low-noise, high-sensitivity detection of light that is imperfectly collimated. While photomultipliers are extraordinarily sensitive and moderately efficient, research is still underway to create a photon-counting light detection device that is much more than 99% efficient.

Such a detector is of interest for applications related to quantum information and quantum cryptography. Elements of photomultiplier technology, when integrated differently, are the basis of night vision devices.

History

Combining two Scientific Discoveries

The invention of the photomultiplier is predicated upon two prior achievements, firstly discovering the photoelectric effect and secondly discovering secondary emission (i.e., the ability of electrons in a vacuum tube to cause the emission of additional electrons by striking an electrode).

Photoelectric Effect

The first demonstration of the photoelectric effect was carried out in 1887 by Heinrich Hertz who demonstrated it using ultraviolet light. Significant for practical applications, Elster and Geitel two years later demonstrated the same effect using *visible* light striking alkali metals (potassium and sodium). The addition of caesium, another alkali metal, has permitted the range of sensitive wavelengths to be extended towards longer wavelengths in the red portion of the visible spectrum.

Historically, the photoelectric effect is associated with Albert Einstein, who relied upon the phenomenon to establish the fundamental principle of quantum mechanics, in 1905, an accomplishment for which Einstein received the 1921 Nobel Prize. It is worthwhile to note that Heinrich Hertz, working 18 years earlier, had not recognized that the kinetic energy of the emitted electrons is proportional to the frequency but independent of the optical intensity. This fact implied a discrete nature of light, i.e. the existence of *quanta*, for the first time.

Secondary Emission

The phenomenon of secondary emission was first limited to purely electronic inventions (i.e., those lacking photosensitivity). In 1902, Austin and Starke reported that the metal surfaces impacted by electron beams emitted a larger number of electrons than were incident. The application of the newly discovered secondary emission to the amplification of signals was only proposed after World War I by Westinghouse scientist Joseph Slepian in a 1919 patent.

The First Photomultiplier

The ingredients for inventing the photomultiplier were coming together during the 1920s as the pace of vacuum tube technologies accelerated. The primary goal for many, if not most, workers was the need for a practical television camera technology. Television had been pursued with primitive prototypes for decades prior to the 1934 introduction of the first practical camera (the iconoscope). Early prototype television cameras lacked sensitivity.

Photomultiplier technology was pursued to enable television camera tubes, such as the iconoscope and (later) the orthicon, to be sensitive enough to be practical. So the stage was set to combine the dual phenomena of photoemission (i.e., the photoelectric effect) with secondary emission, both of which had already been studied and adequately understood, to create a practical photomultiplier.

The first documented photomultiplier demonstration dates to the early 1934 accomplishments of an RCA group based in Harrison, NJ. Harley Iams and Bernard Salzberg were the first to integrate a photoelectric-effect cathode and single secondary emission amplification stage in a single vacuum envelope and the first to characterize its performance as a photomultiplier with electron amplification gain.

These accomplishments were finalized *prior* to June 1934 as detailed in the manuscript submitted to Proceedings of the Institute of Radio Engineers (Proc. IRE). The device consisted of a semi-cylindrical photocathode, a secondary emitter mounted on the axis, and a collector grid surrounding the secondary emitter. The tube had a gain of about eight and operated at frequencies well above 10 kHz.

Higher gains were sought than those available from the early single-stage photomultipliers. However, it is an empirical fact that the yield of secondary electrons is limited in any given secondary emission process, regardless of acceleration voltage. Thus, any single-stage photomultiplier is limited in gain. At the time the maximum first-stage gain that could be achieved was approximately 10 (very significant developments in the 1960s permitted gains above 25 to be reached using negative electron affinity dynodes). For this

reason, multiple-stage photomultipliers, in which the photoelectron yield could be multiplied successively in several stages, were an important goal. The challenge was to cause the photoelectrons to impinge on successively higher-voltage electrodes rather than to travel directly to the highest voltage electrode. Initially this challenge was overcome by using strong magnetic fields to bend the electrons' trajectories. Such a scheme had earlier been conceived by inventor J. Slepian by 1919.

Accordingly, leading international research organizations turned their attention towards improving photomultiplers to achieve higher gain with multiple stages. This work proceeded against a background of economic boom and bust, tyrannical dictatorship, and cataclysmic war clouds collecting on the horizon.

In the USSR, RCA-manufactured radio equipment was introduced on a large scale by Joseph Stalin to construct broadcast networks, and the newly formed All-Union Scientific Research Institute for Television was gearing up a research programme in vacuum tubes that was advanced for its time and place. Numerous visits were made by RCA scientific personnel to the USSR in the 1930s, prior to the Cold War, to instruct the Soviet customers on the capabilities of RCA equipment and to investigate customer needs. During one of these visits, in September 1934, RCA's Vladimir Zworykin was shown the first multiple-dynode photomultiplier, or *photoelectron multiplier*. This pioneering device of 28-year-old Leonid A. Kubetsky achieved gains of 1000x or more when demonstrated in June 1934. The work was submitted for print publication only two years later, in

July 1936 as emphasized in a recent 2006 publication of the Russian Academy of Sciences (RAS). which terms it "Kubetsky's Tube." The Soviet device used a magnetic field to confine the secondary electrons and relied on the Ag-O-Cs photocathode which had been demonstrated by General Electric in the 1920s.

By October 1935, Vladimir Zworykin, George Ashmun Morton, and Louis Malter of RCA in Camden, NJ submitted their manuscript describing the first comprehensive experimental and theoretical analysis of a multiple dynode tube — the device later called a *photomultiplier* — to Proc. IRE. The RCA prototype photomultipliers also used a Ag-O-Cs (silver oxide-caesium) photocathode. They exhibited a peak quantum efficiency of 0.4% at 800 nm.

Whereas these early photomultipliers used the magnetic field principle, electrostatic photomultipliers (with no magnetic field) were demonstrated by Jan Rajchman of RCA Laboratories in Princeton, NJ in the late 1930s and became the standard for all future commercial photomultipliers. The first mass-produced photomultiplier, the Type 931, was of this design and is still commercially produced today.

Also in 1936, a much improved photocathode, Cs_3Sb (caesium-antimony), was reported by P. Gorlich. The caesium-antimony photocathode had a dramatically improved quantum efficiency of 12% at 400 nm, and was used in the first commercially successful photomultipliers manufactured by RCA (i.e., the 931-type) both as a photocathode and as a secondary-emitting material for the dynodes. Different photocathodes provided differing spectral responses.

Spectral Response of Photocathodes

In the early 1940s the JEDEC (Joint Electron Devices Engineering Council), an industry committee on standardization, developed a system of designating spectral responses. The philosophy included the idea that the product's user need only be concerned about the response of the device rather than how the device may be fabricated. Various combinations of photocathode and window materials were assigned "S-numbers" (spectral numbers) ranging from S-1 through S-40, which are still in use today. For example, S-11 uses the caesium-antimony photocathode with a lime glass window, S-13 uses the same photocathode with a fused silica window, and S-25 uses a so-called "multialkali" photocathode (Na-K-Sb-Cs, or sodium-potassium-antimony-caesium) that provides extended response in the red portion of the visible light spectrum. No suitable photoemissive surfaces have yet been reported to detect wavelengths longer than approximately 1700 nanometers, which can be approached by a special (InP/InGaAs(Cs)) photocathode.

Role of RCA

For decades, RCA was responsible for performing the most important work in developing and refining photomultipliers. RCA was also largely responsible for the commercialization of photomultiplers. The company compiled and published an authoritative and very-widely used *Photomultiplier Handbook*. RCA made printed copies available for free upon request. The handbook, which continues to be made available online at no cost by the successors to RCA, is considered to be an essential reference.

Following a corporate break-up in the late 1980s involving the acquisition of RCA by General Electric and disposition of the divisions of RCA to numerous third-parties, RCA's photomultiplier business became an independent company.

Lancaster, Pennsylvania Facility

The Lancaster, Pennsylvania facility was opened by the U.S. Navy in 1942 and operated by RCA for the manufacture of radio and microwave tubes. Following the Allied victory in World War II, the naval facility was acquired by RCA. *RCA Lancaster*, as it became known, was the base for development and production of commercial television products. In subsequent years other products were added, such as cathode ray tubes, photomultiplier tubes, motion-sensing light control switches, and closed-circuit television systems.

Burle Industries

Burle Industries, as a successor to the RCA Corporation, carried the RCA photomultiplier business forward after 1986, based in the Lancaster, Pennsylvania facility. The 1986 acquisition of RCA by General Electric resulted in the divestiture of the RCA Lancaster New Products Division. Hence, 45 years after being founded by the U.S. Navy, its management team, led by Erich Burlefinger, purchased the division and in 1987 founded Burle Industries.

In 2005, after eighteen years as an independent enterprise, Burle Industries and a key subsidiary were acquired by Photonis, a European holding company Photonis Group. Following the acquisition, Photonis was composed of Photonis Netherlands, Photonis France, Photonis USA, and Burle Industries. Photonis USA operates the former Galileo

Corporation Scientific Detector Products Group (Sturbridge, Massachusetts), which had been purchased by Burle Industries in 1999. The Group is known for microchannel plate detector (MCP) electron multipliers—an integrated micro-vacuum tube version of photomultipliers. MCPs are used for imaging and scientific applications, including night vision devices.

On 9 March 2009 Photonis announced that it would cease all production of photomultipliers at both the Lancaster, Pennsylvania and the Brive, France plants.

Other Companies

The Japan-based company Hamamatsu Photonics (also known as Hamamatsu) has emerged since the 1950s as a leader in the photomultiplier industry. Hamamatsu, in the tradition of RCA, has published its own handbook, which is available without cost on the company's website. Hamamatsu uses different designations for particular photocathode formulations and introduces modifications to these designations based on Hamamatsu's proprietary research and development.

Structure and Operating Principles

Photomultipliers are constructed from a glass envelope with a high vacuum inside, which houses a photocathode, several dynodes, and an anode. Incident photons strike the photocathode material, which is present as a thin deposit on the entry window of the device, with electrons being produced as a consequence of the photoelectric effect. These electrons are directed by the focusing electrode toward the electron multiplier, where electrons are multiplied by the

process of secondary emission. The electron multiplier consists of a number of electrodes called *dynodes*. Each dynode is held at a more positive voltage than the previous one. The electrons leave the photocathode, having the energy of the incoming photon (minus the work function of the photocathode).

As the electrons move toward the first dynode, they are accelerated by the electric field and arrive with much greater energy. Upon striking the first dynode, more low energy electrons are emitted, and these electrons in turn are accelerated toward the second dynode. The geometry of the dynode chain is such that a cascade occurs with an ever-increasing number of electrons being produced at each stage. Finally, the electrons reach the anode, where the accumulation of charge results in a sharp current pulse indicating the arrival of a photon at the photocathode. There are two common photomultiplier orientations, the *head-on* or *end-on* (transmission mode) design, as shown above, where light enters the flat, circular top of the tube and passes the photocathode, and the *side-on* design (reflection mode), where light enters at a particular spot on the side of the tube, and impacts on an opaque photocathode. Besides the different photocathode materials, performance is also affected by the transmission of the window material that the light passes through, and by the arrangement of the dynodes. A large number of photomultiplier models are available having various combinations of these, and other, design variables. Either of the manuals mentioned will provide the information needed to choose an appropriate design for a particular application.

Photocathode Materials

The photocathodes can be made of a variety of materials, with different properties. Typically the materials have low work function and are therefore prone to thermionic emission, causing noise and dark current, especially the materials sensitive in infrared; cooling the photocathode lowers this thermal noise. The most common photocathode materials are:

- Ag-O-Cs: also called S1. Transission-mode, sensitive from 300–1200 nm. High dark current; used mainly in near-infrared, with the photocathode cooled.
- GaAs:Cs: caesium-activated gallium arsenide. Flat response from 300 to 850 nm, fading towards ultraviolet and to 930 nm.
- InGaAs:Cs: caesium-activated indium gallium arsenide. Higher infrared sensitivity than GaAs:Cs. Between 900–1000 nm much higher signal-to-noise ratio than Ag-O-Cs.
- Sb-Cs: caesium-activated antimony. Used for reflective mode photocathodes. Response range from ultraviolet to visible. Widely used.
- Bialkali (Sb-K-Cs, Sb-Rb-Cs): caesium-activated antimony-rubidium or antimony-potassium alloy. Similar to Sb:Cs, with higher sensitivity and lower noise. Can be used for transmission-mode; favourable response to a NaI:Tl scintillator flashes makes them widely used in gamma spectroscopy and radiation detection.

  - High-temperature bialkali (Na-K-Sb): can operate up to 175 °C, used in well logging. Low dark current at room temperature.
- Multialkali (Na-K-Sb-Cs): wide spectral response from ultraviolet to near-infrared; special cathode processing can extend range to 930 nm. Used in broadband spectrophotometers.
- Solar-blind (Cs-Te, Cs-I): sensitive to vacuum-UV and ultraviolet. Insensitive to visible light and infrared (CsTe has cutoff at 320 nm, CsI at 200 nm).

Window Materials

The windows of the photomultipliers act as wavelength filters; this may be irrelevant if the cutoff wavelengths are outside of the application range or outside of the photocathode sensitivity range, but special care has to be taken for uncommon wavelengths.

- Borosilicate glass is commonly used for near-infrared to about 300 nm. Glass with very low content of potassium can be used with bialkali photocathodes to lower the background radiation from the potassium-40 isotope.
- Ultraviolet glass transmits visible and ultraviolet down to 185 nm. Used in spectroscopy.
- Synthetic silica transmits down to 160 nm, absorbs less UV than fused silica. Different thermal expansion than kovar (and than borosilicate glass that's expansion-matched to kovar), a graded seal needed between the window and the rest of the tube. The seal is vulnerable to mechanical shocks.

- Magnesium fluoride transmits ultraviolet down to 115 nm. Hygroscopic, though less than other alkali halides usable for UV windows.

Usage Considerations

Photomultiplier tubes typically utilize 1000 to 2000 volts to accelerate electrons within the chain of dynodes. The most negative voltage is connected to the cathode, and the most positive voltage is connected to the anode. Negative high-voltage supplies (with the positive terminal grounded) are preferred, because this configuration enables the photocurrent to be measured at the low voltage side of the circuit for amplification by subsequent electronic circuits operating at low voltage.

Voltages are distributed to the dynodes by a resistive voltage divider, although variations such as active designs (with transistors or diodes) are possible. The divider design, which influences frequency response or rise time, can be selected to suit varying applications. Some instruments that use photomultipliers have provisions to vary the anode voltage to control the gain of the system.

While powered (energized), photomultipliers must be shielded from ambient light to prevent their destruction through overexcitation. If used in a location with strong magnetic fields, which can curve electron paths, steer the electrons away from the dynodes and cause loss of gain, photomultipliers are usually shielded by a layer of mu-metal. This magnetic shield is often maintained at cathode potential. When this is the case, the external shield must also be electrically insulated because of the high voltage on

it. Photomultipliers with large distances between the photocathode and the first dynode are especially sensitive to magnetic fields.

Typical Applications

- Photomultipliers were the first electric eye devices, being used to measure interruptions in beams of light.
- Photomultipliers are used in conjunction with scintillators to detect nuclear and particle radiation in physics experiments.
- Photomultipliers are used in research laboratories to measure the intensity and spectrum of light-emitting materials such as compound semiconductors and quantum dots.
- Photomultipliers are used in numerous medical equipment designs. For example, blood analysis devices used by clinical medical laboratories utilize photomultipliers to determine the relative concentration of various components in vials of blood drawn in doctors' offices, in combination with optical filters and incandescent lamps.

High Sensitivity Applications

After fifty years, during which solid-state electronic components have largely displaced the vacuum tube, the photomultiplier remains a unique and important optoelectronic component. Perhaps its most useful quality is that it acts, electronically, as a nearly perfect current source owing to the high voltage utilized in extracting the tiny currents associated with weak light signals. There is

no Johnson noise associated with photomultiplier signal currents even though they are greatly amplified, e.g., by 100 thousand times (i.e., 100 dB) or more. The photocurrent still contains shot noise. Photomultiplier-amplified photocurrents can be electronically amplified by a high-input-impedance electronic amplifier (in the signal path, subsequent to the photomultiplier), thus producing appreciable voltages even for nearly infinitesimally small photon fluxes. Photomultipliers offer the best possible opportunity to exceed the Johnson noise for many configurations. The aforementioned refers to measurement of light fluxes that, while small, nonetheless amount to a continuous stream of multiple photons.

For smaller photon fluxes, the photomultiplier can be operated in photon counting or Geiger mode. In Geiger mode the photomultiplier gain is set so high (using high voltage) that a single photo-electron resulting from a single photon incident on the primary surface generates a very large current at the output circuit. However, owing to the avalanche of current, a reset of the photomultiplier is required. In either case, the photomultiplier can detect individual photons. The drawback, however, is that not every photon incident on the primary surface is counted either because of less-than-perfect efficiency of the photomultiplier, or because a second photon can arrive at the photomultiplier during the "dead time" associated with a first photon and never be noticed.

A photomultiplier will produce a small current even without incident photons; this is called the *dark current*. Photon counting applications generally demand

photomultipliers designed for low dark current. Nonetheless, the ability to detect single photons striking the primary photosensitive surface itself reveals the quantization principle that Einstein put forth. Photon-counting (as it is called) reveals that light, not only being a wave, consists of discrete particles (i.e., photons).

5

Computer Virus

A computer virus is a computer programme that can copy itself and infect a computer. The term “virus” is also commonly but erroneously used to refer to other types of malware, including but not limited to adware and spyware programmes that do not have the reproductive ability. A true virus can spread from one computer to another (in some form of executable code) when its host is taken to the target computer; for instance because a user sent it over a network or the Internet, or carried it on a removable medium such as a floppy disk, CD, DVD, or USB drive. Viruses can increase their chances of spreading to other computers by infecting files on a network file system or a file system that is accessed by another computer.

As stated above, the term “computer virus” is sometimes used as a catch-all phrase to include all types of malware, even those that do not have the reproductive ability. Malware

includes computer viruses, computer worms, Trojan horses, most rootkits, spyware, dishonest adware and other malicious and unwanted software, including true viruses. Viruses are sometimes confused with worms and Trojan horses, which are technically different. A worm can exploit security vulnerabilities to spread itself automatically to other computers through networks, while a Trojan horse is a programme that appears harmless but hides malicious functions. Worms and Trojan horses, like viruses, may harm a computer system's data or performance. Some viruses and other malware have symptoms noticeable to the computer user, but many are surreptitious or simply do nothing to call attention to themselves. Some viruses do nothing beyond reproducing themselves.

History

Academic Work

The first academic work on the theory of computer viruses (although the term "computer virus" was not invented at that time) was done by John von Neumann in 1949 who held lectures at the University of Illinois about the "Theory and Organization of Complicated Automata". The work of von Neumann was later published as the "Theory of self-reproducing automata". In his essay von Neumann postulated that a computer programme could reproduce. In 1972 Veith Risak published his article "Selbstreproduzierende Automaten mit minimaler Informationsübertragung" (Self-reproducing automata with minimal information exchange). The article describes a fully functional virus written in assembler language for a SIEMENS 4004/35 computer

system. In 1980 Jürgen Kraus wrote his diplom thesis "Selbstreproduktion bei Programmen" (Self-reproduction of programmes) at the University of Dortmund. In his work Kraus postulated that computer programmes can behave in a way similar to biological viruses. In 1984 Fred Cohen from the University of Southern California wrote his paper "Computer Viruses - Theory and Experiments". It was the first paper to explicitly call a self-reproducing programme a "virus"; a term introduced by his mentor Leonard Adleman. An article that describes "useful virus functionalities" was published by J. B. Gunn under the title "Use of virus functions to provide a virtual APL interpreter under user control" in 1984.

Science Fiction

The Terminal Man, a science fiction novel by Michael Crichton (1972), told (as a sideline story) of a computer with telephone modem dialing capability, which had been programmed to randomly dial phone numbers until it hit a modem that is answered by another computer. It then attempted to programme the answering computer with its own programme, so that the second computer would also begin dialing random numbers, in search of yet another computer to programme. The programme is assumed to spread exponentially through susceptible computers. The actual term 'virus' was first used in David Gerrold's 1972 novel, *When HARLIE Was One*. In that novel, a sentient computer named HARLIE writes viral software to retrieve damaging personal information from other computers to blackmail the man who wants to turn him off.

Virus Programmes

The Creeper virus was first detected on ARPANET, the forerunner of the Internet, in the early 1970s. Creeper was an experimental self-replicating programme written by Bob Thomas at BBN Technologies in 1971. Creeper used the ARPANET to infect DEC PDP-10 computers running the TENEX operating system. Creeper gained access via the ARPANET and copied itself to the remote system where the message, "I'm the creeper, catch me if you can!" was displayed. The *Reaper* programme was created to delete Creeper.

A programme called "Elk Cloner" was the first computer virus to appear "in the wild" — that is, outside the single computer or lab where it was created. Written in 1981 by Richard Skrenta, it attached itself to the Apple DOS 3.3 operating system and spread via floppy disk. This virus, created as a practical joke when Skrenta was still in high school, was injected in a game on a floppy disk. On its 50th use the Elk Cloner virus would be activated, infecting the computer and displaying a short poem beginning "Elk Cloner: The programme with a personality."

The first PC virus in the wild was a boot sector virus dubbed (c)Brain, created in 1986 by the Farooq Alvi Brothers in Lahore, Pakistan, reportedly to deter piracy of the software they had written. Before computer networks became widespread, most viruses spread on removable media, particularly floppy disks. In the early days of the personal computer, many users regularly exchanged information and programmes on floppies. Some viruses spread by infecting

programmes stored on these disks, while others installed themselves into the disk boot sector, ensuring that they would be run when the user booted the computer from the disk, usually inadvertently. PCs of the era would attempt to boot first from a floppy if one had been left in the drive. Until floppy disks fell out of use, this was the most successful infection strategy and boot sector viruses were the most common in the wild for many years.

Traditional computer viruses emerged in the 1980s, driven by the spread of personal computers and the resultant increase in BBS, modem use, and software sharing. Bulletin board-driven software sharing contributed directly to the spread of Trojan horse programmes, and viruses were written to infect popularly traded software. Shareware and bootleg software were equally common vectors for viruses on BBS's.

Macro viruses have become common since the mid-1990s. Most of these viruses are written in the scripting languages for Microsoft programmes such as Word and Excel and spread throughout Microsoft Office by infecting documents and spreadsheets. Since Word and Excel were also available for Mac OS, most could also spread to Macintosh computers. Although most of these viruses did not have the ability to send infected email messages, those viruses which did take advantage of the Microsoft Outlook COM interface.

Some old versions of Microsoft Word allow macros to replicate themselves with additional blank lines. If two macro viruses simultaneously infect a document, the combination

of the two, if also self-replicating, can appear as a "mating" of the two and would likely be detected as a virus unique from the "parents".

A virus may also send a web address link as an instant message to all the contacts on an infected machine. If the recipient, thinking the link is from a friend (a trusted source) follows the link to the website, the virus hosted at the site may be able to infect this new computer and continue propagating. Viruses that spread using cross-site scripting were first reported in 2002, and were academically demonstrated in 2005. There have been multiple instances of the cross-site scripting viruses in the wild, exploiting websites such as MySpace and Yahoo.

Infection Strategies

In order to replicate itself, a virus must be permitted to execute code and write to memory. For this reason, many viruses attach themselves to executable files that may be part of legitimate programmes. If a user attempts to launch an infected programme, the virus' code may be executed simultaneously. Viruses can be divided into two types based on their behaviour when they are executed. Nonresident viruses immediately search for other hosts that can be infected, infect those targets, and finally transfer control to the application programme they infected. Resident viruses do not search for hosts when they are started. Instead, a resident virus loads itself into memory on execution and transfers control to the host programme. The virus stays active in the background and infects new hosts when those files are accessed by other programmes or the operating system itself.

Nonresident Viruses

Nonresident viruses can be thought of as consisting of a *finder module* and a *replication module*. The finder module is responsible for finding new files to infect. For each new executable file the finder module encounters, it calls the replication module to infect that file.

Resident Viruses

Resident viruses contain a replication module that is similar to the one that is employed by nonresident viruses. This module, however, is not called by a finder module. The virus loads the replication module into memory when it is executed instead and ensures that this module is executed each time the operating system is called to perform a certain operation. The replication module can be called, for example, each time the operating system executes a file. In this case the virus infects every suitable programme that is executed on the computer.

Resident viruses are sometimes subdivided into a category of *fast infectors* and a category of *slow infectors*. Fast infectors are designed to infect as many files as possible. A fast infector, for instance, can infect every potential host file that is accessed. This poses a special problem when using anti-virus software, since a virus scanner will access every potential host file on a computer when it performs a system-wide scan. If the virus scanner fails to notice that such a virus is present in memory the virus can "piggy-back" on the virus scanner and in this way infect all files that are scanned. Fast infectors rely on their fast infection rate to spread. The disadvantage of this method is that infecting

many files may make detection more likely, because the virus may slow down a computer or perform many suspicious actions that can be noticed by anti-virus software. Slow infectors, on the other hand, are designed to infect hosts infrequently. Some slow infectors, for instance, only infect files when they are copied. Slow infectors are designed to avoid detection by limiting their actions: they are less likely to slow down a computer noticeably and will, at most, infrequently trigger anti-virus software that detects suspicious behaviour by programmes. The slow infector approach, however, does not seem very successful.

Vectors and Hosts

Viruses have targeted various types of transmission media or hosts. This list is not exhaustive:

- Binary executable files (such as COM files and EXE files in MS-DOS, Portable Executable files in Microsoft Windows, the Mach-O format in OSX, and ELF files in Linux)
- Volume Boot Records of floppy disks and hard disk partitions
- The master boot record (MBR) of a hard disk
- General-purpose script files (such as batch files in MS-DOS and Microsoft Windows, VBScript files, and shell script files on Unix-like platforms).
- Application-specific script files (such as Telix-scripts)
- System specific autorun script files (such as Autorun.inf file needed by Windows to automatically run software stored on USB Memory Storage Devices).

- Documents that can contain macros (such as Microsoft Word documents, Microsoft Excel spreadsheets, AmiPro documents, and Microsoft Access database files)
- Cross-site scripting vulnerabilities in web applications
- Arbitrary computer files. An exploitable buffer overflow, format string, race condition or other exploitable bug in a programme which reads the file could be used to trigger the execution of code hidden within it. Most bugs of this type can be made more difficult to exploit in computer architectures with protection features such as an execute disable bit and/or address space layout randomization.

PDFs, like HTML, may *link* to malicious code. PDFs can also be infected with malicious code. In operating systems that use file extensions to determine programme associations (such as Microsoft Windows), the extensions may be hidden from the user by default. This makes it possible to create a file that is of a different type than it appears to the user. For example, an executable may be created named "picture.png.exe", in which the user sees only "picture.png" and therefore assumes that this file is an image and most likely is safe, yet when opened runs the executable on the client machine. An additional method is to generate the virus code from parts of existing operating system files by using the CRC16/CRC32 data. The initial code can be quite small (tens of bytes) and unpack a fairly large virus. This is analogous to a biological "prion" in the way it works but is vulnerable to signature based detection. This attack has not yet been seen "in the wild".

Methods to Avoid Detection

In order to avoid detection by users, some viruses employ different kinds of deception. Some old viruses, especially on the MS-DOS platform, make sure that the "last modified" date of a host file stays the same when the file is infected by the virus. This approach does not fool anti-virus software, however, especially those which maintain and date Cyclic redundancy checks on file changes. Some viruses can infect files without increasing their sizes or damaging the files. They accomplish this by overwriting unused areas of executable files. These are called *cavity viruses*. For example, the CIH virus, or Chernobyl Virus, infects Portable Executable files. Because those files have many empty gaps, the virus, which was 1 KB in length, did not add to the size of the file.

Some viruses try to avoid detection by killing the tasks associated with antivirus software before it can detect them. As computers and operating systems grow larger and more complex, old hiding techniques need to be updated or replaced. Defending a computer against viruses may demand that a file system migrate towards detailed and explicit permission for every kind of file access.

Avoiding Bait Files and Other Undesirable Hosts

A virus needs to infect hosts in order to spread further. In some cases, it might be a bad idea to infect a host programme. For example, many anti-virus programmes perform an integrity check of their own code. Infecting such programmes will therefore increase the likelihood that the virus is detected. For this reason, some viruses are

programmed not to infect programmes that are known to be part of anti-virus software. Another type of host that viruses sometimes avoid are *bait files*. Bait files (or *goat files*) are files that are specially created by anti-virus software, or by anti-virus professionals themselves, to be infected by a virus. These files can be created for various reasons, all of which are related to the detection of the virus:

- Anti-virus professionals can use bait files to take a sample of a virus (i.e. a copy of a programme file that is infected by the virus). It is more practical to store and exchange a small, infected bait file, than to exchange a large application programme that has been infected by the virus.
- Anti-virus professionals can use bait files to study the behaviour of a virus and evaluate detection methods. This is especially useful when the virus is polymorphic. In this case, the virus can be made to infect a large number of bait files. The infected files can be used to test whether a virus scanner detects all versions of the virus.
- Some anti-virus software employs bait files that are accessed regularly. When these files are modified, the anti-virus software warns the user that a virus is probably active on the system.

Since bait files are used to detect the virus, or to make detection possible, a virus can benefit from not infecting them. Viruses typically do this by avoiding suspicious programmes, such as small programme files or programmes that contain certain patterns of 'garbage instructions'. A

related strategy to make baiting difficult is *sparse infection*. Sometimes, sparse infectors do not infect a host file that would be a suitable candidate for infection in other circumstances. For example, a virus can decide on a random basis whether to infect a file or not, or a virus can only infect host files on particular days of the week.

Stealth

Some viruses try to trick antivirus software by intercepting its requests to the operating system. A virus can hide itself by intercepting the antivirus software's request to read the file and passing the request to the virus, instead of the OS. The virus can then return an uninfected version of the file to the antivirus software, so that it seems that the file is "clean". Modern antivirus software employs various techniques to counter stealth mechanisms of viruses. The only completely reliable method to avoid stealth is to boot from a medium that is known to be clean.

Self-modification

Most modern antivirus programmes try to find virus-patterns inside ordinary programmes by scanning them for so-called *virus signatures*. A signature is a characteristic byte-pattern that is part of a certain virus or family of viruses. If a virus scanner finds such a pattern in a file, it notifies the user that the file is infected. The user can then delete, or (in some cases) "clean" or "heal" the infected file. Some viruses employ techniques that make detection by means of signatures difficult but probably not impossible. These viruses modify their code on each infection. That is, each infected file contains a different variant of the virus.

Encryption with a Variable Key

A more advanced method is the use of simple encryption to encipher the virus. In this case, the virus consists of a small decrypting module and an encrypted copy of the virus code. If the virus is encrypted with a different key for each infected file, the only part of the virus that remains constant is the decrypting module, which would (for example) be appended to the end. In this case, a virus scanner cannot directly detect the virus using signatures, but it can still detect the decrypting module, which still makes indirect detection of the virus possible. Since these would be symmetric keys, stored on the infected host, it is in fact entirely possible to decrypt the final virus, but this is probably not required, since self-modifying code is such a rarity that it may be reason for virus scanners to at least flag the file as suspicious. An old, but compact, encryption involves XORing each byte in a virus with a constant, so that the exclusive-or operation had only to be repeated for decryption. It is suspicious for a code to modify itself, so the code to do the encryption/decryption may be part of the signature in many virus definitions.

Polymorphic Code

Polymorphic code was the first technique that posed a serious threat to virus scanners. Just like regular encrypted viruses, a polymorphic virus infects files with an encrypted copy of itself, which is decoded by a decryption module. In the case of polymorphic viruses, however, this decryption module is also modified on each infection. A well-written polymorphic virus therefore has no parts which remain

identical between infections, making it very difficult to detect directly using signatures. Antivirus software can detect it by decrypting the viruses using an emulator, or by statistical pattern analysis of the encrypted virus body.

To enable polymorphic code, the virus has to have a polymorphic engine (also called mutating engine or mutation engine) somewhere in its encrypted body. See Polymorphic code for technical detail on how such engines operate.

Some viruses employ polymorphic code in a way that constrains the mutation rate of the virus significantly. For example, a virus can be programmed to mutate only slightly over time, or it can be programmed to refrain from mutating when it infects a file on a computer that already contains copies of the virus.

The advantage of using such slow polymorphic code is that it makes it more difficult for antivirus professionals to obtain representative samples of the virus, because bait files that are infected in one run will typically contain identical or similar samples of the virus.

This will make it more likely that the detection by the virus scanner will be unreliable, and that some instances of the virus may be able to avoid detection.

Metamorphic Code

To avoid being detected by emulation, some viruses rewrite themselves completely each time they are to infect new executables. Viruses that utilize this technique are said to be metamorphic. To enable metamorphism, a metamorphic engine is needed. A metamorphic virus is usually very large

and complex. For example, W32/Simile consisted of over 14000 lines of Assembly language code, 90% of which is part of the metamorphic engine.

Vulnerability and Countermeasures

The Vulnerability of Operating Systems to Viruses

Just as genetic diversity in a population decreases the chance of a single disease wiping out a population, the diversity of software systems on a network similarly limits the destructive potential of viruses. This became a particular concern in the 1990s, when Microsoft gained market dominance in desktop operating systems and office suites. The users of Microsoft software (especially networking software such as Microsoft Outlook and Internet Explorer) are especially vulnerable to the spread of viruses. Microsoft software is targeted by virus writers due to their desktop dominance, and is often criticized for including many errors and holes for virus writers to exploit. Integrated and non-integrated Microsoft applications (such as Microsoft Office) and applications with scripting languages with access to the file system (for example Visual Basic Script (VBS), and applications with networking features) are also particularly vulnerable. Although Windows is by far the most popular target operating system for virus writers, viruses also exist on other platforms. Any operating system that allows third-party programmes to run can theoretically run viruses. Some operating systems are more secure than others. Unix-based operating systems (and NTFS-aware applications on Windows NT based platforms) only allow their users to run executables within their own protected memory space.

An Internet based experiment revealed that there were cases when people willingly pressed a particular button to download a virus. Security analyst Didier Stevens ran a half year advertising campaign on Google AdWords which said "Is your PC virus-free? Get it infected here!". The result was 409 clicks.

As of 2006, there are relatively few security exploits targeting Mac OS X (with a Unix-based file system and kernel). The number of viruses for the older Apple operating systems, known as Mac OS Classic, varies greatly from source to source, with Apple stating that there are only four known viruses, and independent sources stating there are as many as 63 viruses. Many Mac OS Classic viruses targeted the HyperCard authoring environment. The difference in virus vulnerability between Macs and Windows is a chief selling point, one that Apple uses in their Get a Mac advertising. In January 2009, Symantec announced the discovery of a trojan that targets Macs. This discovery did not gain much coverage until April 2009. While Linux, and Unix in general, has always natively blocked normal users from having access to make changes to the operating system environment, Windows users are generally not. This difference has continued partly due to the widespread use of administrator accounts in contemporary versions like XP. In 1997, when a virus for Linux was released – known as "Bliss" – leading antivirus vendors issued warnings that Unix-like systems could fall prey to viruses just like Windows.

The Bliss virus may be considered characteristic of viruses – as opposed to worms – on Unix systems. Bliss requires that the user run it explicitly, and it can only infect

programmes that the user has the access to modify. Unlike Windows users, most Unix users do not log in as an administrator user except to install or configure software; as a result, even if a user ran the virus, it could not harm their operating system. The Bliss virus never became widespread, and remains chiefly a research curiosity. Its creator later posted the source code to Usenet, allowing researchers to see how it worked.

The Role of Software Development

Because software is often designed with security features to prevent unauthorized use of system resources, many viruses must exploit software bugs in a system or application to spread. Software development strategies that produce large numbers of bugs will generally also produce potential exploits.

Anti-virus Software and Other Preventive Measures

Many users install anti-virus software that can detect and eliminate known viruses after the computer downloads or runs the executable. There are two common methods that an anti-virus software application uses to detect viruses. The first, and by far the most common method of virus detection is using a list of virus signature definitions. This works by examining the content of the computer's memory (its RAM, and boot sectors) and the files stored on fixed or removable drives (hard drives, floppy drives), and comparing those files against a database of known virus "signatures". The disadvantage of this detection method is that users are only protected from viruses that pre-date their last virus definition update. The second method is to use a heuristic

algorithm to find viruses based on common behaviours. This method has the ability to detect novel viruses that anti-virus security firms have yet to create a signature for.

Some anti-virus programmes are able to scan opened files in addition to sent and received email messages "on the fly" in a similar manner. This practice is known as "on-access scanning".

Anti-virus software does not change the underlying capability of host software to transmit viruses. Users must update their software regularly to patch security holes. Anti-virus software also needs to be regularly updated in order to recognize the latest threats.

One may also minimize the damage done by viruses by making regular backups of data (and the operating systems) on different media, that are either kept unconnected to the system (most of the time), read-only or not accessible for other reasons, such as using different file systems.

This way, if data is lost through a virus, one can start again using the backup (which should preferably be recent).

If a backup session on optical media like CD and DVD is closed, it becomes read-only and can no longer be affected by a virus (so long as a virus or infected file was not copied onto the CD/DVD). Likewise, an operating system on a bootable CD can be used to start the computer if the installed operating systems become unusable. Backups on removable media must be carefully inspected before restoration. The Gammima virus, for example, propagates via removable flash drives.

Recovery Methods

Once a computer has been compromised by a virus, it is usually unsafe to continue using the same computer without completely reinstalling the operating system. However, there are a number of recovery options that exist after a computer has a virus. These actions depend on severity of the type of virus.

Virus Removal

One possibility on Windows Me, Windows XP, Windows Vista and Windows 7 is a tool known as System Restore, which restores the registry and critical system files to a previous checkpoint. Often a virus will cause a system to hang, and a subsequent hard reboot will render a system restore point from the same day corrupt.

Restore points from previous days should work provided the virus is not designed to corrupt the restore files or also exists in previous restore points. Some viruses, however, disable System Restore and other important tools such as Task Manager and Command Prompt.

An example of a virus that does this is CiaDoor. However, many such viruses can be removed by rebooting the computer, entering Windows safe mode, and then using system tools.

Administrators have the option to disable such tools from limited users for various reasons (for example, to reduce potential damage from and the spread of viruses). A virus can modify the registry to do the same even if the Administrator is controlling the computer; it blocks *all* users

including the administrator from accessing the tools. The message "Task Manager has been disabled by your administrator" may be displayed, even to the administrator. Users running a Microsoft operating system can access Microsoft's website to run a free scan, provided they have their 20-digit registration number.

Many websites run by anti-virus software companies provide free online virus scanning, with limited cleaning facilities (the purpose of the sites is to sell anti-virus products). Some websites allow a single suspicious file to be checked by many antivirus programmes in one operation.

Operating System Reinstallation

Reinstalling the operating system is another approach to virus removal. It involves either reformatting the computer's hard drive and installing the OS and all programmes from original media, or restoring the entire partition with a clean backup image.

User data can be restored by booting from a Live CD, or putting the hard drive into another computer and booting from its operating system with great care not to infect the second computer by executing any infected programmes on the original drive; and once the system has been restored precautions must be taken to avoid reinfection from a restored executable file.

These methods are simple to do, may be faster than disinfecting a computer, and are guaranteed to remove any malware. If the operating system and programmes must be reinstalled from scratch, the time and effort to reinstall,

reconfigure, and restore user preferences must be taken into account. Restoring from an image is much faster, totally safe, and restores the exact configuration to the state it was in when the image was made, with no further trouble.

6

Computer Antivirus Software

Antivirus or anti-virus software is used to prevent, detect, and remove malware, including but not limited to computer viruses, computer worm, trojan horses, spyware and adware. This page talks about the software used for the prevention and removal of such threats, rather than computer security implemented by software methods. A variety of strategies are typically employed. Signature-based detection involves searching for known patterns of data within executable code. However, it is possible for a computer to be infected with new malware for which no signature is yet known. To counter such so-called zero-day threats, heuristics can be used. One type of heuristic approach, generic signatures, can identify new viruses or variants of existing viruses by looking for known malicious code, or slight variations of such code, in files. Some antivirus software can also predict what a file will do by running it in a sandbox and analyzing

what it does to see if it performs any malicious actions. No matter how useful antivirus software can be, it can sometimes have drawbacks. Antivirus software can impair a computer's performance.

Inexperienced users may also have trouble understanding the prompts and decisions that antivirus software presents them with. An incorrect decision may lead to a security breach. If the antivirus software employs heuristic detection, success depends on achieving the right balance between false positives and false negatives. False positives can be as destructive as false negatives. Finally, antivirus software generally runs at the highly trusted kernel level of the operating system, creating a potential avenue of attack.

History

Most of the computer viruses written in the early and mid 1980s were limited to self-reproduction and had no specific damage routine built into the code. That changed when more and more programmers became acquainted with virus programming and created viruses that manipulated or even destroyed data on infected computers. There are competing claims for the innovator of the first antivirus product.

Possibly the first publicly documented removal of a computer virus in the wild was performed by Bernd Fix in 1987. Fred Cohen, who published one of the first academic papers on computer viruses in 1984, began to develop strategies for antivirus software in 1988 that were picked up and continued by later antivirus software developers.

Also in 1988 a mailing list named VIRUS-L was started on the BITNET/EARN network where new viruses and the possibilities of detecting and eliminating viruses were discussed.

Some members of this mailing list like John McAfee or Eugene Kaspersky later founded software companies that developed and sold commercial antivirus software. Before internet connectivity was widespread, viruses were typically spread by infected floppy disks. Antivirus software came into use, but was updated relatively infrequently. During this time, virus checkers essentially had to check executable files and the boot sectors of floppy disks and hard disks. However, as internet usage became common, viruses began to spread online. Over the years it has become necessary for antivirus software to check an increasing variety of files, rather than just executables, for several reasons:

- Powerful macros used in word processor applications, such as Microsoft Word, presented a risk. Virus writers could use the macros to write viruses embedded within documents. This meant that computers could now also be at risk from infection by opening documents with hidden attached macros.
- Later email programmes, in particular Microsoft's Outlook Express and Outlook, were vulnerable to viruses embedded in the email body itself. A user's computer could be infected by just opening or previewing a message.

As always-on broadband connections became the norm, and more and more viruses were released, it became essential

to update virus checkers more and more frequently. Even then, a new zero-day virus could become widespread before antivirus companies released an update to protect against it.

Firmware Issues

Active anti-virus software can interfere with a firmware update process. Any writeable firmware in the computer can be infected by malicious code. This is a major concern, as an infected BIOS could require the actual BIOS chip to be replaced to ensure the malicious code is completely removed. Anti-virus software is not effective at protecting firmware and the motherboard BIOS from infection.

Other Methods

Installed antivirus software running on an individual computer is only one method of guarding against viruses. Other methods are also used, including cloud-based antivirus, firewalls and on-line scanners.

Cloud Antivirus

Cloud antivirus is a technology that uses lightweight agent software on the protected computer, while offloading the majority of data analysis to the provider's infrastructure. One approach to implementing cloud antivirus involves scanning suspicious files using multiple antivirus engines. This approach was proposed by an early implementation of the cloud antivirus concept called CloudAV. CloudAV was designed to send programmes or documents to a network cloud where multiple antivirus and behavioural detection programmes are used simultaneously in order to improve

detection rates. Parallel scanning of files using potentially incompatible antivirus scanners is achieved by spawning a virtual machine per detection engine and therefore eliminating any possible issues. CloudAV can also perform "retrospective detection," whereby the cloud detection engine rescans all files in its file access history when a new threat is identified thus improving new threat detection speed. Finally, CloudAV is a solution for effective virus scanning on devices that lack the computing power to perform the scans themselves.

Network Firewall

Network firewalls prevent unknown programmes and processes from accessing the system. However, they are not antivirus systems and make no attempt to identify or remove anything. They may protect against infection from outside the protected computer or network, and limit the activity of any malicious software which is present by blocking incoming or outgoing requests on certain TCP/IP ports. A firewall is designed to deal with broader system threats that come from network connections into the system and is not an alternative to a virus protection system.

Online Scanning

Some antivirus vendors maintain websites with free online scanning capability of the entire computer, critical areas only, local disks, folders or files. Periodic online scanning is a good idea for those than run antivirus applications on their computers because those apps are frequently slow to catch threats. One of the first things that malicious software does in an attack is disable any existing antivirus software

and sometimes the only way to know of an attack is by turning to an online resource that isn't already installed on the infected computer.

Specialist Tools

Virus removal tools are available to help remove stubborn infections or certain types of infection. Examples include Trend Micro's *Rootkit Buster*, and rkhunter for the detection of rootkits, Avira's *AntiVir Removal Tool*, *PCTools Threat Removal Tool*, and AVG's Anti-Virus Free 2011. A rescue disk that is bootable, such as a CD or USB storage device, can be used to run antivirus software outside of the installed operating system, in order to remove infections while they are dormant. A bootable antivirus disk can be useful when, for example, the installed operating system is no longer bootable or has malware that is resisting all attempts to be removed by the installed antivirus software. Examples of some of these bootable disks include the *Avira AntiVir Rescue System*, *PCTools Alternate Operating System Scanner*, and *AVG Rescue CD*. The AVG Rescue CD software can also be installed onto a USB storage device, that is bootable on newer computers.

Popularity

A survey by Symantec in 2009 found that a third of small to medium sized business did not use antivirus protection at that time, whereas more than 80% of home users had some kind of antivirus installed.

Identification Methods

There are several methods which antivirus software can use to identify malware. Signature based detection is the

most common method. To identify viruses and other malware, antivirus software compares the contents of a file to a dictionary of virus signatures. Because viruses can embed themselves in existing files, the entire file is searched, not just as a whole, but also in pieces. Heuristic-based detection, like malicious activity detection, can be used to identify unknown viruses. File emulation is another heuristic approach.

File emulation involves executing a programme in a virtual environment and logging what actions the programme performs. Depending on the actions logged, the antivirus software can determine if the programme is malicious or not and then carry out the appropriate disinfection actions.

Signature Based Detection

Traditionally, antivirus software heavily relied upon signatures to identify malware. This can be very effective, but cannot defend against malware unless samples have already been obtained and signatures created. Because of this, signature-based approaches are not effective against new, unknown viruses. As new viruses are being created each day, the signature-based detection approach requires frequent updates of the virus signature dictionary. To assist the antivirus software companies, the software may allow the user to upload new viruses or variants to the company, allowing the virus to be analyzed and the signature added to the dictionary. Although the signature-based approach can effectively contain virus outbreaks, virus authors have tried to stay a step ahead of such software by writing "oligomorphic", "polymorphic" and, more recently,

"metamorphic" viruses, which encrypt parts of themselves or otherwise modify themselves as a method of disguise, so as to not match virus signatures in the dictionary.

Heuristics

Some more sophisticated antivirus software uses heuristic analysis to identify new malware or variants of known malware. Many viruses start as a single infection and through either mutation or refinements by other attackers, can grow into dozens of slightly different strains, called variants. Generic detection refers to the detection and removal of multiple threats using a single virus definition. For example, the Vundo trojan has several family members, depending on the antivirus vendor's classification. Symantec classifies members of the Vundo family into two distinct categories, *Trojan.Vundo* and *Trojan.Vundo.B*. While it may be advantageous to identify a specific virus, it can be quicker to detect a virus family through a generic signature or through an inexact match to an existing signature. Virus researchers find common areas that all viruses in a family share uniquely and can thus create a single generic signature. These signatures often contain non-contiguous code, using wildcard characters where differences lie. These wildcards allow the scanner to detect viruses even if they are padded with extra, meaningless code. A detection that uses this method is said to be "heuristic detection."

Rootkit Detection

Anti-virus software can also scan for rootkits; a rootkit is a type of malware that is designed to gain administrative-level control over a computer system without being detected.

Rootkits can change how the operating system functions and in some cases can tamper with the anti-virus programme and render it ineffective. Rootkits are also difficult to remove, in some cases requiring a complete re-installation of the operating system.

Issues of Concern

Unexpected Renewal Costs

Some commercial antivirus software end-user license agreements include a clause that the subscription will be automatically renewed, and the purchaser's credit card automatically billed, at the renewal time without explicit approval. For example, McAfee requires users to unsubscribe at least 60 days before the expiration of the present subscription while BitDefender sends notifications to unsubscribe 30 days before the renewal. Norton Antivirus also renews subscriptions automatically by default.

Rogue Security Applications

Some apparent antivirus programmes are actually malware masquerading as legitimate software, such as WinFixer and MS Antivirus.

Problems Caused by False Positives

A "false positive" is when antivirus software identifies a non-malicious file as a virus. When this happens, it can cause serious problems. For example, if an antivirus programme is configured to immediately delete or quarantine infected files, a false positive in a essential file can render the operating system or some applications unusable. In May 2007, a faulty virus signature issued by Symantec

mistakenly removed essential operating system files, leaving thousands of PCs unable to boot. Also in May 2007 the executable file required by Pegasus Mail was falsely detected by Norton AntiVirus as being a Trojan and it was automatically removed, preventing Pegasus Mail from running. Norton anti-virus has falsely identified three releases of Pegasus Mail as malware, and would delete the Pegasus Mail installer file when this happens. In response to this Pegasus Mail stated:

> "On the basis that Norton/Symantec has done this for every one of the last three releases of Pegasus Mail, we can only condemn this product as too flawed to use, and recommend in the strongest terms that our users cease using it in favour of alternative, less buggy anti-virus packages."

In April 2010 McAfee VirusScan detected svchost.exe, a normal Windows binary, as a virus on machines running Windows XP with Service Pack 3, causing a reboot loop and loss of all network access. In December 2010, a faulty update on the AVG anti-virus suite damaged 64-bit versions of Windows 7, rendering it unable to boot, due to an endless boot loop created. When Microsoft Windows becomes damaged by faulty anti-virus products, fixing the damage to Microsoft Windows incurs technical support costs and businesses can be forced to close whilst remedial action is undertaken.

System and Interoperability Related Issues

Running multiple antivirus programmes concurrently can degrade performance and create conflicts. However,

using a concept called multiscanning, several companies (including G Data and Microsoft) have created applications which can run multiple engines concurrently. It is sometimes necessary to temporarily disable virus protection when installing major updates such as Windows Service Packs or updating graphics card drivers.

Active antivirus protection may partially or completely prevent the installation of a major update. A minority of software programmes are not compatible with anti-virus software. For example, the TrueCrypt troubleshooting page reports that anti-virus programmes can conflict with TrueCrypt and cause it to malfunction. Support issues also exist around antivirus application interoperability with common solutions like SSL VPN remote access and network access control products. These technology solutions often have policy assessment applications which require that an up to date antivirus is installed and running. If the antivirus application is not recognized by the policy assessment, whether because the antivirus application has been updated or because it is not part of the policy assessment library, the user will be unable to connect.

Effectiveness

Studies in December 2007 showed that the effectiveness of antivirus software had decreased in the previous year, particularly against unknown or zero day attacks. The computer magazine *c't* found that detection rates for these threats had dropped from 40-50% in 2006 to 20-30% in 2007. At that time, the only exception was the NOD32 antivirus, which managed a detection rate of 68 percent.

The problem is magnified by the changing intent of virus authors. Some years ago it was obvious when a virus infection was present. The viruses of the day, written by amateurs, exhibited destructive behaviour or pop-ups. Modern viruses are often written by professionals, financed by criminal organizations. Independent testing on all the major virus scanners consistently shows that none provide 100% virus detection. The best ones provided as high as 99.6% detection, while the lowest provided only 81.8% in tests conducted in February 2010. All virus scanners produce false positive results as well, identifying benign files as malware. Although methodologies may differ, some notable independent quality testing agencies include AV-Comparatives, ICSA Labs, West Coast Labs, VB100 and other members of the Anti-Malware Testing Standards Organization.

New Viruses

Anti-virus programmes are not always effective against new viruses, even those that use non-signature-based methods that should detect new viruses. The reason for this is that the virus designers test their new viruses on the major anti-virus applications to make sure that they are not detected before releasing them into the wild. Some new viruses, particularly ransomware, use polymorphic code to avoid detection by virus scanners. Jerome Segura, a security analyst with ParetoLogic, explained:

> "It's something that they miss a lot of the time because this type of [ransomware virus] comes from sites that use a polymorphism, which means they basically randomize the file they send you

> and it gets by well-known antivirus products very easily. I've seen people firsthand getting infected, having all the pop-ups and yet they have antivirus software running and it's not detecting anything. It actually can be pretty hard to get rid of, as well, and you're never really sure if it's really gone. When we see something like that usually we advise to reinstall the operating system or reinstall backups."

A proof of concept virus has used the Graphics Processing Unit (GPU) to avoid detection from anti-virus software. The potential success of this involves bypassing the CPU in order to make it much harder for security researchers to analyse the inner workings of such malware.

Rootkits

Detecting rootkits is a major challenge for anti-virus programmes. Rootkits have full administrative access to the computer and are invisible to users and hidden from the list of running processes in the task manager. Rootkits can modify the inner workings of the operating system and tamper with antivirus programmes.

Damaged Files

Files which have been damaged by computer viruses are normally damaged beyond recovery. Anti-virus software removes the virus code from the file during disinfection, but this does not always restore the file to its undamaged state. In such circumstances, damaged files can only be restored from existing backups.

7

Malware Antivirus Software

Malware, short for *malicious software*, (sometimes referred to as pestware) is a software designed to harm or secretly access a computer system without the owner's informed consent. The expression is a general term used by computer professionals to mean a variety of forms of hostile, intrusive, or annoying software or programme code. Software is considered to be malware based on the perceived intent of the creator rather than any particular features.

Malware includes computer viruses, worms, trojan horses, spyware, dishonest adware, scareware, crimeware, most rootkits, and other malicious and unwanted software or programme. In law, malware is sometimes known as a computer contaminant, for instance in the legal codes of several U.S. states, including California and West Virginia. Preliminary results from Symantec published in 2008 suggested that "the release rate of malicious code and other

unwanted programmes may be exceeding that of legitimate software applications." According to F-Secure, "As much malware [was] produced in 2007 as in the previous 20 years altogether."

Malware's most common pathway from criminals to users is through the Internet: primarily by e-mail and the World Wide Web.

The prevalence of malware as a vehicle for organized Internet crime, along with the general inability of traditional anti-malware protection platforms (products) to protect against the continuous stream of unique and newly produced malware, has seen the adoption of a new mindset for businesses operating on the Internet: the acknowledgment that some sizable percentage of Internet customers will always be infected for some reason or another, and that they need to continue doing business with infected customers.

The result is a greater emphasis on back-office systems designed to spot fraudulent activities associated with advanced malware operating on customers' computers.

On March 29, 2010, Symantec Corporation named Shaoxing, China, as the world's malware capital. Malware is not the same as defective software, that is, software that has a legitimate purpose but contains harmful bugs. Sometimes, malware is disguised as genuine software, and may come from an official site. Therefore, some security programmes, such as McAfee may call malware "potentially unwanted programmes" or "PUP". Though a computer virus is malware that can reproduce itself, the term is often used erroneously to refer to the entire category.

Purposes

Many early infectious programmes, including the first Internet Worm and a number of MS-DOS viruses, were written as experiments or pranks. They were generally intended to be harmless or merely annoying, rather than to cause serious damage to computer systems.

In some cases, the perpetrator did not realize how much harm his or her creations would do. Young programmers learning about viruses and their techniques wrote them simply for practice, or to see how far they could spread. As late as 1999, widespread viruses such as the Melissa virus and the David virus appear to have been written chiefly as pranks. The first mobile phone virus, Cabir, appeared in 2004.

Hostile intent related to vandalism can be found in programmes designed to cause harm or data loss. Many DOS viruses, and the Windows ExploreZip worm, were designed to destroy files on a hard disk, or to corrupt the file system by writing invalid data to them. Network-borne worms such as the 2001 Code Red worm or the Ramen worm fall into the same category.

Designed to vandalize web pages, worms may seem like the online equivalent to graffiti tagging, with the author's alias or affinity group appearing everywhere the worm goes. Since the rise of widespread broadband Internet access, malicious software has been designed for a profit, for examples forced advertising. For instance, since 2003, the majority of widespread viruses and worms have been designed to take control of users' computers for black-

market exploitation. Infected "zombie computers" are used to send email spam, to host contraband data such as child pornography, or to engage in distributed denial-of-service attacks as a form of extortion. Another strictly for-profit category of malware has emerged in spyware — programmes designed to monitor users' web browsing, display unsolicited advertisements, or redirect affiliate marketing revenues to the spyware creator.

Spyware programmes do not spread like viruses; they are, in general, installed by exploiting security holes or are packaged with user-installed software, such as peer-to-peer applications.

Infectious Malware: Viruses and Worms

The best-known types of malware, *viruses* and *worms*, are known for the manner in which they spread, rather than any other particular behaviour. The term *computer virus* is used for a programme that has infected some executable software and, *when run*, causes the virus to spread to other executables. Viruses may also contain a payload that performs other actions, often malicious. On the other hand, a *worm* is a programme that *actively* transmits itself over a network to infect other computers. It too may carry a payload.

These definitions lead to the observation that a virus requires user intervention to spread, whereas a worm spreads itself automatically. Using this distinction, infections transmitted by email or Microsoft Word documents, which rely on the recipient opening a file or email to infect the system, would be classified as viruses rather than worms.

Some writers in the trade and popular press misunderstand this distinction and use the terms interchangeably.

Capsule History of Viruses and Worms

Before Internet access became widespread, viruses spread on personal computers by infecting the executable boot sectors of floppy disks. By inserting a copy of itself into the machine code instructions in these executables, a virus causes itself to be run whenever a programme is run or the disk is booted. Early computer viruses were written for the Apple II and Macintosh, but they became more widespread with the dominance of the IBM PC and MS-DOS system. Executable-infecting viruses are dependent on users exchanging software or boot-able floppies, so they spread rapidly in computer hobbyist circles.

The first worms, network-borne infectious programmes, originated not on personal computers, but on multitasking Unix systems. The first well-known worm was the Internet Worm of 1988, which infected SunOS and VAX BSD systems. Unlike a virus, this worm did not insert itself into other programmes. Instead, it exploited security holes (vulnerabilities) in network server programmes and started itself running as a separate process. This same behaviour is used by today's worms as well.

With the rise of the Microsoft Windows platform in the 1990s, and the flexible macros of its applications, it became possible to write infectious code in the macro language of Microsoft Word and similar programmes. These *macro viruses* infect documents and templates rather than applications (executables), but rely on the fact that macros in a Word

document are a form of executable code. Today, worms are most commonly written for the Windows OS, although a few like Mare-D and the Lion worm are also written for Linux and Unix systems. Worms today work in the same basic way as 1988's Internet Worm: they scan the network and leverage vulnerable computers to replicate. Because they need no human intervention, worms can spread with incredible speed. The SQL Slammer infected thousands of computers in a few minutes.

Characteristics of Data-stealing Malware

Does not leave traces of the event

- The malware is typically stored in a cache that is routinely flushed
- The malware may be installed via a drive-by-download process
- The website hosting the malware as well as the malware is generally temporary or rogue

Frequently changes and extends its functions

- It is difficult for antivirus software to detect final payload attributes due to the combination(s) of malware components
- The malware uses multiple file encryption levels

Thwarts Intrusion Detection Systems (IDS) after successful installation

- There are no perceivable network anomalies
- The malware hides in web traffic
- The malware is stealthier in terms of traffic and resource use

Thwarts Disk Encryption

- Data is stolen during decryption and display
- The malware can record keystrokes, passwords, and screenshots

Thwarts Data Loss Prevention (DLP)

- Leakage protection hinges on metadata tagging, not everything is tagged
- Miscreants can use encryption to port data

Examples of Data-stealing Malware

- Bancos, an info stealer that waits for the user to access banking websites then spoofs pages of the bank website to steal sensitive information.
- Gator, spyware that covertly monitors web-surfing habits, uploads data to a server for analysis then serves targeted pop-up ads.
- LegMir, spyware that steals personal information such as account names and passwords related to online games.
- Qhost, a Trojan that modifies the Hosts file to point to a different DNS server when banking sites are accessed then opens a spoofed login page to steal login credentials for those financial institutions.

Data-stealing Malware Incidents

- Albert Gonzalez (not to be confused with the U.S. Attorney General Alberto Gonzalez) is accused of masterminding a ring to use malware to steal and sell more than 170 million credit card numbers in 2006 and 2007—the largest computer fraud in history.

Among the firms targeted were BJ's Wholesale Club, TJX, DSW Shoe, OfficeMax, Barnes & Noble, Boston Market, Sports Authority and Forever 21.

- A Trojan horse programme stole more than 1.6 million records belonging to several hundred thousand people from Monster Worldwide Inc's job search service. The data was used by cybercriminals to craft phishing emails targeted at Monster.com users to plant additional malware on users' PCs.
- Customers of Hannaford Bros. Co, a supermarket chain based in Maine, were victims of a data security breach involving the potential compromise of 4.2 million debit and credit cards. The company was hit by several class-action law suits.
- The Torpig Trojan has compromised and stolen login credentials from approximately 250,000 online bank accounts as well as a similar number of credit and debit cards. Other information such as email, and FTP accounts from numerous websites, have also been compromised and stolen.

Controversy About Assignment to Spyware

There is a group of software (Alexa toolbar, Google toolbar, Eclipse data usage collector, etc) that send data to a central server about which pages have been visited or which features of the software have been used. However differently from "classic" malware these tools document activities and only send data with the user's approval. The user may opt in to share the data in exchange to the additional features and services, or (in case of Eclipse) as the form of voluntary

support for the project. Some security tools report such loggers as malware while others do not. The status of the group is questionable. Some tools like PDFCreator are more on the boundary than others because opting out has been made more complex than it could be (during the installation, the user needs to uncheck two check boxes rather than one). However also PDFCreator is only sometimes mentioned as malware and is still subject of discussions.

Vulnerability to Malware

In this context, as throughout, it should be borne in mind that the "system" under attack may be of various types, e.g. a single computer and operating system, a network or an application. Various factors make a system more vulnerable to malware:

- Homogeneity: e.g. when all computers in a network run the same OS, upon exploiting one, one can exploit them all.
- Weight of numbers: simply because the vast majority of existing malware is written to attack Windows systems, then Windows systems, ipso facto, are more vulnerable to succumbing to malware (regardless of the security strengths or weaknesses of Windows itself).
- Defects: malware leveraging defects in the OS design.
- Unconfirmed code: code from a floppy disk, CD-ROM or USB device may be executed without the user's agreement.
- Over-privileged users: some systems allow all users to modify their internal structures.

- Over-privileged code: some systems allow code executed by a user to access all rights of that user.

An oft-cited cause of vulnerability of networks is homogeneity or software monoculture. For example, Microsoft Windows or Apple Mac have such a large share of the market that concentrating on either could enable a cracker to subvert a large number of systems, but any total monoculture is a problem.

Instead, introducing inhomogeneity (diversity), purely for the sake of robustness, could increase short-term costs for training and maintenance. However, having a few diverse nodes would deter total shutdown of the network, and allow those nodes to help with recovery of the infected nodes. Such separate, functional redundancy would avoid the cost of a total shutdown, would avoid homogeneity as the problem of "all eggs in one basket". Most systems contain bugs, or loopholes, which may be exploited by malware.

A typical example is the buffer-overrun weakness, in which an interface designed to store data, in a small area of memory, allows the caller to supply more data than will fit. This extra data then overwrites the interface's own executable structure (past the end of the buffer and other data).

In this manner, malware can force the system to execute malicious code, by replacing legitimate code with its own payload of instructions (or data values) copied into live memory, outside the buffer area. Originally, PCs had to be booted from floppy disks, and until recently it was common for this to be the default boot device.

This meant that a corrupt floppy disk could subvert the computer during booting, and the same applies to CDs. Although that is now less common, it is still possible to forget that one has changed the default, and rare that a BIOS makes one confirm a boot from removable media. In some systems, non-administrator users are over-privileged by design, in the sense that they are allowed to modify internal structures of the system.

In some environments, users are over-privileged because they have been inappropriately granted administrator or equivalent status. This is primarily a configuration decision, but on Microsoft Windows systems the default configuration is to over-privilege the user. This situation exists due to decisions made by Microsoft to prioritize compatibility with older systems above security configuration in newer systems and because typical applications were developed without the under-privileged users in mind. As privilege escalation exploits have increased this priority is shifting for the release of Microsoft Windows Vista. As a result, many existing applications that require excess privilege (over-privileged code) may have compatibility problems with Vista. However, Vista's User Account Control feature attempts to remedy applications not designed for under-privileged users, acting as a crutch to resolve the privileged access problem inherent in legacy applications.

Malware, running as over-privileged code, can use this privilege to subvert the system. Almost all currently popular operating systems, and also many scripting applications allow code too many privileges, usually in the sense that

when a user executes code, the system allows that code all rights of that user. This makes users vulnerable to malware in the form of e-mail attachments, which may or may not be disguised.

Given this state of affairs, users are warned only to open attachments they trust, and to be wary of code received from untrusted sources. It is also common for operating systems to be designed so that device drivers need escalated privileges, while they are supplied by more and more hardware manufacturers.

Eliminating Over-privileged Code

Over-privileged code dates from the time when most programmes were either delivered with a computer or written in-house, and repairing it would at a stroke render most antivirus software almost redundant. It would, however, have appreciable consequences for the user interface and system management. The system would have to maintain privilege profiles, and know which to apply for each user and programme. In the case of newly installed software, an administrator would need to set up default profiles for the new code. Eliminating vulnerability to rogue device drivers is probably harder than for arbitrary rogue executables. Two techniques, used in VMS, that can help are memory mapping only the registers of the device in question and a system interface associating the driver with interrupts from the device.

Other approaches are:

- Various forms of virtualization, allowing the code unlimited access only to virtual resources

- Various forms of sandbox or jail
- The security functions of Java, in java.security

Such approaches, however, if not fully integrated with the operating system, would reduplicate effort and not be universally applied, both of which would be detrimental to security.

Anti-malware Programmes

As malware attacks become more frequent, attention has begun to shift from viruses and spyware protection, to malware protection, and programmes have been developed to specifically combat them. Anti-malware programmes can combat malware in two ways:

1. They can provide real time protection against the installation of malware software on a computer. This type of spyware protection works the same way as that of antivirus protection in that the anti-malware software scans all incoming network data for malware software and blocks any threats it comes across.
2. Anti-malware software programmes can be used solely for detection and removal of malware software that has already been installed onto a computer. This type of malware protection is normally much easier to use and more popular. This type of anti-malware software scans the contents of the Windows registry, operating system files, and installed programmes on a computer and will provide a list of any threats found, allowing the user to choose which files to delete or keep, or to compare this list to a list of known malware components, removing files that match.

Real-time protection from malware works identically to real-time antivirus protection: the software scans disk files at download time, and blocks the activity of components known to represent malware. In some cases, it may also intercept attempts to install start-up items or to modify browser settings. Because many malware components are installed as a result of browser exploits or user error, using security software (some of which are anti-malware, though many are not) to "sandbox" browsers (essentially babysit the user and their browser) can also be effective in helping to restrict any damage done.

Academic Research on Malware: A Brief Overview

The notion of a self-reproducing computer programme can be traced back to when presented lectures that encompassed the theory and organization of complicated automata. Neumann showed that in theory a programme could reproduce itself. This constituted a plausibility result in computability theory. Fred Cohen experimented with computer viruses and confirmed Neumann's postulate. He also investigated other properties of malware (detectability, self-obfuscating programmes that used rudimentary encryption that he called "evolutionary", and so on). His 1988 doctoral dissertation was on the subject of computer viruses. Cohen's faculty advisor, Leonard Adleman (the A in RSA) presented a rigorous proof that, in the general case, algorithmically determining whether a virus is or is not present is Turing undecidable. This problem must not be mistaken for that of determining, within a broad class of programmes, that a virus is not present; this problem differs in that it does not require the ability to recognize

all viruses. Adleman's proof is perhaps the deepest result in malware computability theory to date and it relies on Cantor's diagonal argument as well as the halting problem. Ironically, it was later shown by Young and Yung that Adleman's work in cryptography is ideal in constructing a virus that is highly resistant to reverse-engineering by presenting the notion of a cryptovirus.

A cryptovirus is a virus that contains and uses a public key and randomly generated symmetric cipher initialization vector (IV) and session key (SK). In the cryptoviral extortion attack, the virus hybrid encrypts plaintext data on the victim's machine using the randomly generated IV and SK.

The IV+SK are then encrypted using the virus writer's public key. In theory the victim must negotiate with the virus writer to get the IV+SK back in order to decrypt the ciphertext (assuming there are no backups). Analysis of the virus reveals the public key, not the IV and SK needed for decryption, or the private key needed to recover the IV and SK. This result was the first to show that computational complexity theory can be used to devise malware that is robust against reverse-engineering.

Another growing area of computer virus research is to mathematically model the infection behaviour of worms using models such as Lotka–Volterra equations, which has been applied in the study of biological virus. Various virus propagation scenarios have been studied by researchers such as propagation of computer virus, fighting virus with virus like predator codes, effectiveness of patching etc.

Grayware

Grayware (or Greynet) is a general term sometimes used as a classification for applications that behave in a manner that is annoying or undesirable, and yet less serious or troublesome than malware.

Grayware encompasses spyware, adware, dialers, joke programmes, remote access tools, and any other unwelcome files and programmes apart from viruses that are designed to harm the performance of computers on your network. The term has been in use since at least as early as September 2004.

Grayware refers to applications or files that are not classified as viruses or trojan horse programmes, but can still negatively affect the performance of the computers on your network and introduce significant security risks to your organization.

Often grayware performs a variety of undesired actions such as irritating users with pop-up windows, tracking user habits and unnecessarily exposing computer vulnerabilities to attack.

- Spyware is software that installs components on a computer for the purpose of recording Web surfing habits (primarily for marketing purposes). Spyware sends this information to its author or to other interested parties when the computer is online. Spyware often downloads with items identified as 'free downloads' and does not notify the user of its existence or ask for permission to install the components. The information spyware components

gather can include user keystrokes, which means that private information such as login names, passwords, and credit card numbers are vulnerable to theft.

- Adware is software that displays advertising banners on Web browsers such as Internet Explorer and Mozilla Firefox. While not categorized as malware, many users consider adware invasive. Adware programmes often create unwanted effects on a system, such as annoying popup ads and the general degradation in either network connection or system performance. Adware programmes are typically installed as separate programmes that are bundled with certain free software. Many users inadvertently agree to installing adware by accepting the End User License Agreement (EULA) on the free software. Adware are also often installed in tandem with spyware programmes. Both programmes feed off each other's functionalities: spyware programmes profile users' Internet behaviour, while adware programmes display targeted ads that correspond to the gathered user profile.

Web and Spam

The World Wide Web is a criminals' preferred pathway for spreading malware.

Today's web threats use combinations of malware to create infection chains. About one in ten Web pages may contain malicious code.

Wikis and Blogs

Attackers may use wikis and blogs to advertise links that lead to malware sites. Wiki and blog servers can also be attacked directly. In 2010, Network Solutions was compromised and some hosted sites became a path to malware and spam.

Targeted SMTP Threats

Targeted SMTP threats also represent an emerging attack vector through which malware is propagated. As users adapt to widespread spam attacks, cybercriminals distribute crimeware to target one specific organization or industry, often for financial gain.

HTTP and FTP

Infections via "drive-by" download are spread through the Web over HTTP and FTP when resources containing spurious keywords are indexed by legitimate search engines, as well as when JavaScript is surreptitiously added to legitimate websites and advertising networks.

Concealment: Trojan Horses, Rootkits, and Backdoors

Trojan Horses

For a malicious programme to accomplish its goals, it must be able to run without being shut down, or deleted by the user or administrator of the computer system on which it is running. Concealment can also help get the malware installed in the first place. When a malicious programme is disguised as something innocuous or desirable, users may be tempted to install it without knowing what

it does. This is the technique of the *Trojan horse* or *trojan*. In broad terms, a Trojan horse is any programme that invites the user to run it, concealing a harmful or malicious payload.

The payload may take effect immediately and can lead to many undesirable effects, such as deleting the user's files or further installing malicious or undesirable software. Trojan horses known as droppers are used to start off a worm outbreak, by injecting the worm into users' local networks. One of the most common ways that spyware is distributed is as a Trojan horse, bundled with a piece of desirable software that the user downloads from the Internet. When the user installs the software, the spyware is installed alongside. Spyware authors who attempt to act in a legal fashion may include an end-user license agreement that states the behaviour of the spyware in loose terms, which the users are unlikely to read or understand.

Rootkits

Once a malicious programme is installed on a system, it is essential that it *stays* concealed, to avoid detection and disinfection. The same is true when a human attacker breaks into a computer directly. Techniques known as *rootkits* allow this concealment, by modifying the host's operating system so that the malware is hidden from the user. Rootkits can prevent a malicious process from being visible in the system's list of processes, or keep its files from being read. Originally, a rootkit was a set of tools installed by a human attacker on a Unix system, allowing the attacker to gain administrator (root) access. Today, the term is used

more generally for concealment routines in a malicious programme. Some malicious programmes contain routines to defend against removal, not merely to hide themselves, but to repel attempts to remove them. An early example of this behaviour is recorded in the Jargon File tale of a pair of programmes infesting a Xerox CP-V time sharing system: Each ghost-job would detect the fact that the other had been killed, and would start a new copy of the recently slain programme within a few milliseconds. The only way to kill both ghosts was to kill them simultaneously (very difficult) or to deliberately crash the system.

Similar techniques are used by some modern malware, wherein the malware starts a number of processes that monitor and restore one another as needed. In the event a user running Microsoft Windows is infected with such malware, if they wish to manually stop it, they could use Task Manager's 'processes' tab to find the main process (the one that spawned the "resurrector process(es)"), and use the 'end process tree' function, which would kill not only the main process, but the "resurrector(s)" as well, since they were started by the main process.

Some malware programmes use other techniques, such as naming the infected file similar to a legitimate or trustworthy file (expl0rer.exe VS explorer.exe).

Backdoors

A backdoor is a method of bypassing normal authentication procedures. Once a system has been compromised (by one of the above methods, or in some other way), one or more backdoors may be installed in order

to allow easier access in the future. Backdoors may also be installed prior to malicious software, to allow attackers entry. The idea has often been suggested that computer manufacturers preinstall backdoors on their systems to provide technical support for customers, but this has never been reliably verified. Crackers typically use backdoors to secure remote access to a computer, while attempting to remain hidden from casual inspection. To install backdoors crackers may use Trojan horses, worms, or other methods.

Malware for Profit: Spyware, Botnets, Keystroke Loggers, and Dialers

During the 1980s and 1990s, it was usually taken for granted that malicious programmes were created as a form of vandalism or prank. More recently, the greater share of malware programmes have been written with a profit motive (financial or otherwise) in mind. This can be taken as the malware authors' choice to monetize their control over infected systems: to turn that control into a source of revenue. Spyware programmes are commercially produced for the purpose of gathering information about computer users, showing them pop-up ads, or altering web-browser behaviour for the financial benefit of the spyware creator. For instance, some spyware programmes redirect search engine results to paid advertisements. Others, often called "stealware" by the media, overwrite affiliate marketing codes so that revenue is redirected to the spyware creator rather than the intended recipient.

Spyware programmes are sometimes installed as Trojan horses of one sort or another. They differ in that their

creators present themselves openly as businesses, for instance by selling advertising space on the pop-ups created by the malware. Most such programmes present the user with an end-user license agreement that purportedly protects the creator from prosecution under computer contaminant laws. However, spyware EULAs have not yet been upheld in court.

Another way that financially motivated malware creators can profit from their infections is to directly use the infected computers to do work for the creator. The infected computers are used as proxies to send out spam messages. A computer left in this state is often known as a zombie computer. The advantage to spammers of using infected computers is they provide anonymity, protecting the spammer from prosecution. Spammers have also used infected PCs to target anti-spam organizations with distributed denial-of-service attacks.

In order to coordinate the activity of many infected computers, attackers have used coordinating systems known as *botnets*. In a botnet, the malware or malbot logs in to an Internet Relay Chat channel or other chat system. The attacker can then give instructions to all the infected systems simultaneously. Botnets can also be used to push upgraded malware to the infected systems, keeping them resistant to antivirus software or other security measures.

It is possible for a malware creator to profit by stealing sensitive information from a victim. Some malware programmes install a *key logger*, which intercepts the user's keystrokes when entering a password, credit card number, or other information that may be exploited. This is then

transmitted to the malware creator automatically, enabling credit card fraud and other theft. Similarly, malware may copy the CD key or password for online games, allowing the creator to steal accounts or virtual items. Another way of stealing money from the infected PC owner is to take control of a dial-up modem and dial an expensive toll call. *Dialer* (or *porn dialer*) software dials up a premium-rate telephone number such as a U.S. "900 number" and leave the line open, charging the toll to the infected user.

Data-stealing Malware

Data-stealing malware is a web threat that divests victims of personal and proprietary information with the intent of monetizing stolen data through direct use or underground distribution. Content security threats that fall under this umbrella include keyloggers, screen scrapers, spyware, adware, backdoors, and bots. The term does not refer to activities such as spam, phishing, DNS poisoning, SEO abuse, etc. However, when these threats result in file download or direct installation, as most hybrid attacks do, files that act as agents to proxy information will fall into the data-stealing malware category.

8

Computer User Uses of Spyware

Spyware is a type of malware that can be installed on computers, and which collects small pieces of information about users without their knowledge. The presence of spyware is typically hidden from the user, and can be difficult to detect. Typically, spyware is secretly installed on the user's personal computer. Sometimes, however, spywares such as keyloggers are installed by the owner of a shared, corporate, or public computer on purpose in order to secretly monitor other users.

While the term *spyware* suggests software that secretly monitors the user's computing, the functions of spyware extend well beyond simple monitoring. Spyware programmes can collect various types of personal information, such as Internet surfing habits and sites that have been visited, but can also interfere with user control of the computer in other ways, such as installing additional software and redirecting

Web browser activity. Spyware is known to change computer settings, resulting in slow connection speeds, different home pages, and/or loss of Internet connection or functionality of other programmes. In an attempt to increase the understanding of spyware, a more formal classification of its included software types is provided by the term privacy-invasive software.

In response to the emergence of spyware, a small industry has sprung up dealing in anti-spyware software. Running anti-spyware software has become a widely recognized element of computer security practices for computers, especially those running Microsoft Windows. A number of jurisdictions have passed anti-spyware laws, which usually target any software that is surreptitiously installed to control a user's computer.

History and Development

The first recorded use of the term spyware occurred on 16 October 1995 in a Usenet post that poked fun at Microsoft's business model. *Spyware* at first denoted *software* meant for espionage purposes. However, in early 2000 the founder of Zone Labs, Gregor Freund, used the term in a press release for the ZoneAlarm Personal Firewall. Since then, "spyware" has taken on its present sense. According to a 2005 study by AOL and the National Cyber-Security Alliance, 61 percent of surveyed users' computers were infected with form of spyware. 92 percent of surveyed users with spyware reported that they did not know of its presence, and 91 percent reported that they had not given permission for the installation of the spyware. As of 2006, spyware has

become one of the preeminent security threats to computer systems running Microsoft Windows operating systems. Computers on which Internet Explorer (IE) is the primary browser are particularly vulnerable to such attacks, not only because IE is the most widely-used, but because its tight integration with Windows allows spyware access to crucial parts of the operating system.

Before Internet Explorer 6 SP2 was released as part of Windows XP Service Pack 2, the browser would automatically display an installation window for any ActiveX component that a website wanted to install.

The combination of user naivety concerning malware, and the assumption by Internet Explorer that all ActiveX components are benign, led, in part, to the massive spread of spyware. Many spyware components would also make use of exploits in Javascript, Internet Explorer and Windows to install without user knowledge or permission. The Windows Registry contains multiple sections where modification of key values allows software to be executed automatically when the operating system boots.

Spyware can exploit this design to circumvent attempts at removal. The spyware typically will link itself from each location in the registry that allows execution. Once running, the spyware will periodically check if any of these links are removed. If so, they will be automatically restored. This ensures that the spyware will execute when the operating system is booted, even if some (or most) of the registry links are removed.

Digital Rights Management

Some copy-protection technologies have borrowed from spyware. In 2005, Sony BMG Music Entertainment was found to be using rootkits in its XCP digital rights management technology Like spyware, not only was it difficult to detect and uninstall, it was so poorly written that most efforts to remove it could have rendered computers unable to function.

Texas Attorney General Greg Abbott filed suit, and three separate class-action suits were filed. Sony BMG later provided a workaround on its website to help users remove it. Beginning on 25 April 2006, Microsoft's Windows Genuine Advantage Notifications application was installed on most Windows PCs as a "critical security update". While the main purpose of this deliberately uninstallable application is to ensure the copy of Windows on the machine was lawfully purchased and installed, it also installs software that has been accused of "phoning home" on a daily basis, like spyware. It can be removed with the RemoveWGA tool.

Personal Relationships

Spyware has been used to surreptitiously monitor electronic activities of partners in intimate relationships, generally to uncover evidence of infidelity. At least one software package, Loverspy, was specifically marketed for this purpose. Depending on local laws regarding communal/ marital property, observing a partner's online activity without their consent may be illegal; the author of Loverspy and several users of the product were indicted in California in 2005 on charges of wiretapping and various computer crimes.

Browser Cookies

Anti-spyware programmes often report Web advertisers' HTTP cookies, the small text files that track browsing activity, as spyware. While they are not always inherently malicious, many users object to third parties using space on their personal computers for their business purposes, and many anti-spyware programmes offer to remove them.

Examples

These common spyware programmes illustrate the diversity of behaviours found in these attacks. Note that as with computer viruses, researchers give names to spyware programmes which may not be used by their creators. Programmes may be grouped into "families" based not on shared programme code, but on common behaviours, or by "following the money" of apparent financial or business connections. For instance, a number of the spyware programmes distributed by Claria are collectively known as "Gator". Likewise, programmes that are frequently installed together may be described as parts of the same spyware package, even if they function separately.

- CoolWebSearch, a group of programmes, takes advantage of Internet Explorer vulnerabilities. The package directs traffic to advertisements on Web sites including *coolwebsearch.com*. It displays pop-up ads, rewrites search engine results, and alters the infected computer's hosts file to direct DNS lookups to these sites.
- Internet Optimizer, also known as DyFuCa, redirects Internet Explorer error pages to advertising. When

users follow a broken link or enter an erroneous URL, they see a page of advertisements. However, because password-protected Web sites (HTTP Basic authentication) use the same mechanism as HTTP errors, Internet Optimizer makes it impossible for the user to access password-protected sites.

- HuntBar, aka WinTools or Adware.Websearch, was installed by an ActiveX drive-by download at affiliate Web sites, or by advertisements displayed by other spyware programmes—an example of how spyware can install more spyware. These programmes add toolbars to IE, track aggregate browsing behaviour, redirect affiliate references, and display advertisements.
- Movieland, also known as Moviepass.tv and Popcorn.net, is a movie download service that has been the subject of thousands of complaints to the Federal Trade Commission (FTC), the Washington State Attorney General's Office, the Better Business Bureau, and other agencies. Consumers complained they were held hostage by a cycle of oversized pop-up windows demanding payment of at least $29.95, claiming that they had signed up for a three-day free trial but had not cancelled before the trial period was over, and were thus obligated to pay. The FTC filed a complaint, since settled, against Movieland and eleven other defendants charging them with having "engaged in a nationwide scheme to use deception and coercion to extract payments from consumers."

- WeatherStudio has a plugin that displays a window-panel near the *bottom* of a browser window. The official website notes that it is easy to remove (uninstall) WeatherStudio from a computer, using its own uninstall-programme, such as under C:\Programme Files\WeatherStudio. Once WeatherStudio is removed, a browser returns to the prior display appearance, without the need to modify the browser settings.
- Zango (formerly 180 Solutions) transmits detailed information to advertisers about the Web sites which users visit. It also alters HTTP requests for affiliate advertisements linked from a Web site, so that the advertisements make unearned profit for the 180 Solutions company. It opens pop-up ads that cover over the Web sites of competing companies (as seen in their [Zango End User License Agreement]).
- Zlob trojan, or just Zlob, downloads itself to a computer via an ActiveX codec and reports information back to Control Server. Some information can be the search-history, the Websites visited, and even keystrokes. More recently, Zlob has been known to hijack routers set to defaults.

Comparison

Spyware, Adware and Tracking

The term *adware* frequently refers to any software which displays advertisements, whether or not the user has consented. Programmes such as the Eudora mail client display advertisements as an alternative to shareware

registration fees. These may be classified as "adware", in the sense of advertising-supported software, but not as spyware. Adware in this form does not operate surreptitiously or mislead the user, and provides the user with a specific service.

Most adware is *spyware* in a different sense than "advertising-supported software": it displays advertisements related to what it finds from spying on users. Gator Software from Claria Corporation (formerly GATOR) and Exact Advertising's BargainBuddy are examples. Visited Web sites frequently install Gator on client machines in a surreptitious manner, and it directs revenue to the installing site and to Claria by displaying advertisements to the user. The user is shown many pop-up advertisements.

Other spyware behaviour, such as reporting on websites the user visits, occurs in the background. The data is used for "targeted" advertisement impressions. The prevalence of spyware has cast suspicion on other programmes that track Web browsing, even for statistical or research purposes.

Some observers describe the Alexa Toolbar, an Internet Explorer plug-in published by Amazon.com, as spyware, and some anti-spyware programmes such as Ad-Aware report it as such. Many of these adware-distributing companies are backed by millions of dollars of adware-generating revenues. Adware and spyware are similar to viruses in that they can be considered malicious in nature. People are profiting from misleading adware, sometimes known as scareware, such as Antivirus 2009.

Similarly, software bundled with free, advertising-supported programmes such as P2P acts as spyware (and, if removed, disables the 'parent' programme), yet people are willing to download it. This presents a dilemma for proprietors of anti-spyware products whose removal tools may inadvertently disable wanted programmes. For example, WhenUSave is ignored by popular anti-spyware programme Ad-Aware (but removed as spyware by most scanners) because it is part of the popular (but recently decommissioned) eDonkey client. To address this dilemma, the Anti-Spyware Coalition was formed to establish and document best practices regarding acceptable software behaviour.

Spyware, Viruses and Worms

Unlike viruses and worms, spyware does not usually self-replicate. Like many recent viruses, however, spyware—by design—exploits infected computers for commercial gain. Typical tactics include delivery of unsolicited pop-up advertisements, theft of personal information (including financial information such as credit card numbers), monitoring of Web-browsing activity for marketing purposes, and routing of HTTP requests to advertising sites. However, spyware can be dropped as a payload by a worm.

Routes of Infection

Spyware does not directly spread in the manner of a computer virus or worm: generally, an infected system does not attempt to transmit the infection to other computers. Instead, spyware gets on a system through deception of the user or through exploitation of software vulnerabilities.

Most spyware is installed without users' knowledge. Since they tend not to install software if they know that it will disrupt their working environment and compromise their privacy, spyware deceives users, either by piggybacking on a piece of desirable software such as Kazaa, or by tricking them into installing it (the Trojan horse method). Some "rogue" spyware programmes masquerade as security software.

The distributor of spyware usually presents the programme as a useful utility—for instance as a "Web accelerator" or as a helpful software agent. Users download and install the software without immediately suspecting that it could cause harm. For example, Bonzi Buddy, a programme bundled with spyware and targeted at children, claims that:

> *He will explore the Internet with you as your very own friend and sidekick! He can talk, walk, joke, browse, search, e-mail, and download like no other friend you've ever had! He even has the ability to compare prices on the products you love and help you save money! Best of all, he's FREE!*

Spyware can also come bundled with other software. The user downloads a programme and installs it, and the installer additionally installs the spyware. Although the desirable software itself may do no harm, the bundled spyware does. In some cases, spyware authors have paid shareware authors to bundle spyware with their software. In other cases, spyware authors have repackaged desirable freeware with installers that slipstream spyware.

Some spyware authors infect a system through security holes in the Web browser or in other software. When the user navigates to a Web page controlled by the spyware author, the page contains code which attacks the browser and forces the download and installation of spyware. The spyware author would also have some extensive knowledge of commercially-available anti-virus and firewall software. This has become known as a "drive-by download", which leaves the user a hapless bystander to the attack. Common browser exploits target security vulnerabilities in Internet Explorer and in the Sun Microsystems Java runtime.

The installation of spyware frequently involves Internet Explorer. Its popularity and history of security issues have made it the most frequent target. Its deep integration with the Windows environment and scriptability make it an obvious point of attack into Windows. Internet Explorer also serves as a point of attachment for spyware in the form of Browser Helper Objects, which modify the browser's behaviour to add toolbars or to redirect traffic.

In a few cases, a worm or virus has delivered a spyware payload. Some attackers used the Spybot worm to install spyware that put pornographic pop-ups on the infected system's screen. By directing traffic to ads set up to channel funds to the spyware authors, they profit personally.

Effects and Behaviours

A spyware programme is rarely alone on a computer: an affected machine usually has multiple infections. Users frequently notice unwanted behaviour and degradation of system performance. A spyware infestation can create

significant unwanted CPU activity, disk usage, and network traffic. Stability issues, such as applications freezing, failure to boot, and system-wide crashes, are also common. Spyware, which interferes with networking software, commonly causes difficulty connecting to the Internet.

In some infections, the spyware is not even evident. Users assume in those situations that the performance issues relate to faulty hardware, Windows installation problems, or another infection. Some owners of badly infected systems resort to contacting technical support experts, or even buying a new computer because the existing system "has become too slow". Badly infected systems may require a clean reinstallation of all their software in order to return to full functionality.

Only rarely does a single piece of software render a computer unusable. Rather, a computer is likely to have multiple infections. The cumulative effect, and the interactions between spyware components, causes the symptoms commonly reported by users: a computer, which slows to a crawl, overwhelmed by the many parasitic processes running on it.

Moreover, some types of spyware disable software firewalls and anti-virus software, and/or reduce browser security settings, thus opening the system to further opportunistic infections, much like an immune deficiency disease. Some spyware disables or even removes competing spyware programmes, on the grounds that more spyware-related annoyances make it even more likely that users will take action to remove the programmes. One spyware maker,

Avenue Media, even sued a competitor, Direct Revenue, over this; the two later settled with an agreement not to disable each others' products. Some other types of spyware use rootkit like techniques to prevent detection, and thus removal. Targetsoft, for instance, modifies the "Winsock" Windows Sockets files. The deletion of the spyware-infected file "inetadpt.dll" will interrupt normal networking usage. A typical Windows user has administrative privileges, mostly for convenience. Because of this, any programme the user runs (intentionally or not) has unrestricted access to the system.

As with other operating systems, Windows users too are able to follow the principle of least privilege and use non-administrator least user access accounts, or to reduce the privileges of specific vulnerable Internet-facing processes such as Internet Explorer (through the use of tools such as DropMyRights).

However, as this is not a default configuration, few users do this. In Windows Vista, by default, a computer administrator runs everything under limited user privileges. When a programme requires administrative privileges, Vista will prompt the user with an allow/deny pop-up. This improves on the design used by previous versions of Windows.

Advertisements

Many spyware programmes display advertisements. Some programmes simply display pop-up ads on a regular basis; for instance, one every several minutes, or one when the user opens a new browser window. Others display ads in

response to the user visiting specific sites. Spyware operators present this feature as desirable to advertisers, who may buy ad placement in pop-ups displayed when the user visits a particular site. It is also one of the purposes for which spyware programmes gather information on user behaviour.

Many users complain about irritating or offensive advertisements as well. As with many banner ads, spyware advertisements often use animation or flickering banners, which can be visually distracting and annoying to users. Pop-up ads for pornography often display indiscriminately. Links to these sites may be added to the browser window, history or search function.

When children are the users, this could possibly violate anti-pornography laws in some jurisdictions. A number of spyware programmes break the boundaries of illegality; variations of "Zlob.Trojan" and "Trojan-Downloader. Win32.INService" have been known to show undesirable child pornography, key gens, cracks and illegal software pop-up ads, which violate child pornography and copyright laws. A further issue in the case of some spyware programmes concerns the replacement of banner ads on viewed web sites. Spyware that acts as a web proxy or a Browser Helper Object can replace references to a site's own advertisements (which fund the site) with advertisements that instead fund the spyware operator. This cuts into the margins of advertising-funded Web sites.

"Stealware" and Affiliate Fraud

A few spyware vendors, notably 180 Solutions, have written what the *New York Times* has dubbed "stealware",

and what spyware researcher Ben Edelman terms *affiliate fraud*, a form of click fraud. Stealware diverts the payment of affiliate marketing revenues from the legitimate affiliate to the spyware vendor.

Spyware which attacks affiliate networks places the spyware operator's affiliate tag on the user's activity — replacing any other tag, if there is one. The spyware operator is the only party that gains from this. The user has their choices thwarted, a legitimate affiliate loses revenue, networks' reputations are injured, and vendors are harmed by having to pay out affiliate revenues to an "affiliate" who is not party to a contract. Affiliate fraud is a violation of the terms of service of most affiliate marketing networks. As a result, spyware operators such as 180 Solutions have been terminated from affiliate networks including LinkShare and ShareSale.

Identity Theft and Fraud

In one case, spyware has been closely associated with identity theft. In August 2005, researchers from security software firm Sunbelt Software suspected the creators of the common CoolWebSearch spyware had used it to transmit "chat sessions, user names, passwords, bank information, etc."; however it turned out that "it actually (was) its own sophisticated criminal little trojan that's independent of CWS." This case is currently under investigation by the FBI. The Federal Trade Commission estimates that 27.3 million Americans have been victims of identity theft, and that financial losses from identity theft totaled nearly $48 billion for businesses and financial institutions and at least $5 billion in out-of-pocket expenses for individuals.

Spyware-makers may commit wire fraud with *dialer* programme spyware. These can reset a modem to dial up a premium-rate telephone number instead of the usual ISP. Connecting to these suspicious numbers involves long-distance or overseas charges which invariably result in high call costs. Dialers are ineffective on computers that do not have a modem, or are not connected to a telephone line, and are now very rare due to the decline in use of dial-up internet access.

Legal Issues

Criminal Law

Unauthorized access to a computer is illegal under computer crime laws, such as the U.S. Computer Fraud and Abuse Act, the U.K.'s Computer Misuse Act, and similar laws in other countries. Since owners of computers infected with spyware generally claim that they never authorized the installation, a *prima facie* reading would suggest that the promulgation of spyware would count as a criminal act. Law enforcement has often pursued the authors of other malware, particularly viruses. However, few spyware developers have been prosecuted, and many operate openly as strictly legitimate businesses, though some have faced lawsuits. Spyware producers argue that, contrary to the users' claims, users do in fact give consent to installations. Spyware that comes bundled with shareware applications may be described in the legalese text of an end-user license agreement (EULA). Many users habitually ignore these purported contracts, but spyware companies such as Claria say these demonstrate that users have consented.

Despite the ubiquity of EULAs and of "clickwrap" agreements, under which a single click can be taken as consent to the entire text, relatively little caselaw has resulted from their use. It has been established in most common law jurisdictions that a clickwrap agreement can be a binding contract *in certain circumstances.* This does not, however, mean that every such agreement is a contract, or that every term in one is enforceable.

Some jurisdictions, including the U.S. states of Iowa and Washington, have passed laws criminalizing some forms of spyware. Such laws make it illegal for anyone other than the owner or operator of a computer to install software that alters Web-browser settings, monitors keystrokes, or disables computer-security software. In the United States, lawmakers introduced a bill in 2005 entitled the Internet Spyware Prevention Act, which would imprison creators of spyware.

Administrative Sanctions

US FTC Actions

The US Federal Trade Commission has sued Internet marketing organizations under the "unfairness doctrine" to make them stop infecting consumers' PCs with spyware. In one case, that against Seismic Entertainment Productions, the FTC accused the defendants of developing a programme that seized control of PCs nationwide, infected them with spyware and other malicious software, bombarded them with a barrage of pop-up advertising for Seismic's clients, exposed the PCs to security risks, and caused them to malfunction, slow down, and, at times, crash. Seismic then offered to sell the victims an "antispyware" programme to

fix the computers, and stop the popups and other problems that Seismic had caused. On November 21, 2006, a settlement was entered in federal court under which a $1.75 million judgment was imposed in one case and $1.86 million in another, but the defendants were insolvent

In a second case, brought against CyberSpy Software LLC, the FTC charged that CyberSpy marketed and sold "RemoteSpy" keylogger spyware to clients who would then secretly monitor unsuspecting consumers' computers. According to the FTC, Cyberspy touted RemoteSpy as a "100% undetectable" way to "Spy on Anyone. From Anywhere." The FTC has obtained a temporary order prohibiting the defendants from selling the software and disconnecting from the Internet any of their servers that collect, store, or provide access to information that this software has gathered. The case is still in its preliminary stages. A complaint filed by the Electronic Privacy Information Center (EPIC) brought the RemoteSpy software to the FTC's attention.

Netherlands OPTA

An administrative fine, the first of its kind in Europe, has been issued by the Independent Authority of Posts and Telecommunications (OPTA) from the Netherlands. It applied fines in total value of Euro 1,000,000 for infecting 22 million computers. The spyware concerned is called DollarRevenue. The law articles that have been violated are art. 4.1 of the Decision on universal service providers and on the interests of end users; the fines have been issued based on art. 15.4 taken together with art. 15.10 of the Dutch

telecommunications law. A part of these fines has to be paid personally by the directors of these companies, i.e. not from the accounts of their companies, but from their personal fortunes. Since an appeal has been lodged, the fines will have to be paid only after a Dutch law court makes a decision in this case. The culprits maintain that the evidence for violating the two law articles has been obtained illegally. The names of the directors and the names of the companies have not been revealed, since it is not clear that OPTA is allowed to make such information public.

Civil Law

Former New York State Attorney General and former Governor of New York Eliot Spitzer has pursued spyware companies for fraudulent installation of software. In a suit brought in 2005 by Spitzer, the California firm Intermix Media, Inc. ended up settling, by agreeing to pay US$7.5 million and to stop distributing spyware. The hijacking of Web advertisements has also led to litigation. In June 2002, a number of large Web publishers sued Claria for replacing advertisements, but settled out of court. Courts have not yet had to decide whether advertisers can be held liable for spyware that displays their ads. In many cases, the companies whose advertisements appear in spyware pop-ups do not directly do business with the spyware firm. Rather, they have contracted with an advertising agency, which in turn contracts with an online subcontractor who gets paid by the number of “impressions” or appearances of the advertisement. Some major firms such as Dell Computer and Mercedes-Benz have sacked advertising agencies that have run their ads in spyware.

Libel Suits by Spyware Developers

Litigation has gone both ways. Since "spyware" has become a common pejorative, some makers have filed libel and defamation actions when their products have been so described. In 2003, Gator (now known as Claria) filed suit against the website PC Pitstop for describing its programme as "spyware". PC Pitstop settled, agreeing not to use the word "spyware", but continues to describe harm caused by the Gator/Claria software. As a result, other anti-spyware and anti-virus companies have also used other terms such as "potentially unwanted programmes" or greyware to denote these products.

WebcamGate

In the 2010 WebcamGate case, plaintiffs charged two suburban Philadelphia high schools secretly spied on students by surreptitiously and remotely activating webcams embedded in school-issued laptops the students were using at home, and therefore infringed on their privacy rights. The school loaded each student's computer with LANrev's remote activation tracking software. This included the now-discontinued "TheftTrack". While TheftTrack was not enabled by default on the software, the programme allowed the school district to elect to activate it, and to choose which of the TheftTrack surveillance options the school wanted to enable.

TheftTrack allowed school district employees to secretly remotely activate a tiny webcam embedded in the student's laptop, above the laptop's screen. That allowed school officials to secretly take photos through the webcam, of whatever

was in front of it and in its line of sight, and send the photos to the school's server. The LANrev software disabled the webcams for all other uses (*e.g.*, students were unable to use Photo Booth or video chat), so most students mistakenly believed their webcams did not work at all. In addition to webcam surveillance, TheftTrack allowed school officials to take screenshots, and send them to the school's server. In addition, LANrev allowed school officials to take snapshots of instant messages, web browsing, music playlists, and written compositions. The schools admitted to secretly snapping over 66,000 webshots and screenshots, including webcam shots of students in their bedrooms.

Remedies and Prevention

As the spyware threat has worsened, a number of techniques have emerged to counteract it. These include programmes designed to remove or to block spyware, as well as various user practices which reduce the chance of getting spyware on a system. Nonetheless, spyware remains a costly problem. When a large number of pieces of spyware have infected a Windows computer, the only remedy may involve backing up user data, and fully reinstalling the operating system. For instance, some versions of Vundo cannot be completely removed by Symantec, Microsoft, PC Tools, and others because it infects rootkit, Internet Explorer, and Windows' lsass.exe (Local Security Authority Subsystem Service) with a randomly-filenamed dll (dynamic link library).

Anti-spyware Programmes

Many programmers and some commercial firms have released products dedicated to remove or block spyware.

Steve Gibson's *OptOut* pioneered a growing category. Programmes such as PC Tools' Spyware Doctor, Lavasoft's *Ad-Aware SE* (free scans for non-commercial users, must pay for other features) and Patrick Kolla's *Spybot - Search & Destroy* (all features free for non-commercial use) rapidly gained popularity as effective tools to remove, and in some cases intercept, spyware programmes. On December 16, 2004, Microsoft acquired the *GIANT AntiSpyware* software, rebranding it as *Windows AntiSpyware beta* and releasing it as a free download for Genuine Windows XP and Windows 2003 users. In 2006, Microsoft renamed the beta software to Windows Defender (free), and it was released as a free download in October 2006 and is included as standard with Windows Vista as well as Windows 7. Major anti-virus firms such as Symantec, PC Tools, McAfee and Sophos have come later to the table, adding anti-spyware features to their existing anti-virus products. Early on, anti-virus firms expressed reluctance to add anti-spyware functions, citing lawsuits brought by spyware authors against the authors of web sites and programmes which described their products as "spyware". However, recent versions of these major firms' home and business anti-virus products do include anti-spyware functions, albeit treated differently from viruses. Symantec Anti-Virus, for instance, categorizes spyware programmes as "extended threats" and now offers real-time protection from them (as it does for viruses).

Recently, the anti-virus company Grisoft, creator of AVG Anti-Virus, acquired anti-spyware firm Ewido Networks, re-labeling their Ewido anti-spyware programme as AVG Anti-Spyware Professional Edition. AVG also used this product

to add an integrated anti-spyware solution to some versions of the AVG Anti-Virus family of products, and a freeware AVG Anti-Spyware Free Edition available for private and non-commercial use. This shows a trend by anti virus companies to launch a dedicated solution to spyware and malware. Zone Labs, creator of Zone Alarm firewall have also released an anti-spyware programme.

Anti-spyware programmes can combat spyware in two ways:

1. They can provide real time protection against the installation of spyware software on the computer. This type of spyware protection works the same way as that of anti-virus protection in that the anti-spyware software scans all incoming network data for spyware software and blocks any threats it comes across.
2. Anti-spyware software programmes can be used solely for detection and removal of spyware software that has already been installed onto the computer. This type of spyware protection is normally much easier to use and more popular. With this spyware protection software the user can schedule weekly, daily, or monthly scans of the computer to detect and remove any spyware software that have been installed on the computer. This type of anti-spyware software scans the contents of the windows registry, operating system files, and installed programmes on the computer and will provide a list of any threats found, allowing the user to choose what to delete and what to keep.

Such programmes inspect the contents of the Windows registry, the operating system files, and installed programmes, and remove files and entries which match a list of known spyware components. Real-time protection from spyware works identically to real-time anti-virus protection: the software scans disk files at download time, and blocks the activity of components known to represent spyware.

In some cases, it may also intercept attempts to install start-up items or to modify browser settings. Because many spyware and adware are installed as a result of browser exploits or user error, using security software (some of which are antispyware, though many are not) to sandbox browsers can also be effective to help restrict any damage done. Earlier versions of anti-spyware programmes focused chiefly on detection and removal. Javacool Software's SpywareBlaster, one of the first to offer real-time protection, blocked the installation of ActiveX-based and other spyware programmes.

Like most anti-virus software, many anti-spyware/adware tools require a frequently-updated database of threats. As new spyware programmes are released, anti-spyware developers discover and evaluate them, making "signatures" or "definitions" which allow the software to detect and remove the spyware. As a result, anti-spyware software is of limited usefulness without a regular source of updates. Some vendors provide a subscription-based update service, while others provide updates free. Updates may be installed automatically on a schedule or before doing a scan, or may be done manually.

Not all programmes rely on updated definitions. Some programmes rely partly (for instance many antispyware programmes such as Windows Defender, Spybot's TeaTimer and Spysweeper) or fully (programmes falling under the class of HIPS such as BillP's WinPatrol) on historical observation. They watch certain configuration parameters (such as certain portions of the Windows registry or browser configuration) and report any change to the user, without judgment or recommendation.

While they do not rely on updated definitions, which may allow them to spot newer spyware, they can offer no guidance. The user is left to determine "what did I just do, and is this configuration change appropriate?" Windows Defender's SpyNet attempts to alleviate this through offering a community to share information, which helps guide both users, who can look at decisions made by others, and analysts, who can spot fast-spreading spyware. A popular generic spyware removal tool used by those with a certain degree of expertise is HijackThis, which scans certain areas of the Windows OS where spyware often resides and presents a list with items to delete manually. As most of the items are legitimate windows files/registry entries it is advised for those who are less knowledgeable on this subject to post a HijackThis log on the numerous antispyware sites and let the experts decide what to delete.

If a spyware programme is not blocked and manages to get itself installed, it may resist attempts to terminate or uninstall it. Some programmes work in pairs: when an anti-spyware scanner (or the user) terminates one running process, the other one respawns the killed programme.

Likewise, some spyware will detect attempts to remove registry keys and immediately add them again. Usually, booting the infected computer in safe mode allows an anti-spyware programme a better chance of removing persistent spyware. Killing the process tree may also work.

A new breed of spyware (Look2Me spyware by NicTechNetworks is a good example) hides inside system-critical processes and start up even in safe mode, see rootkit. With no process to terminate they are harder to detect and remove. Sometimes they do not even leave any on-disk signatures. Rootkit technology is also seeing increasing use, as is the use of NTFS alternate data streams. Newer spyware programmes also have specific countermeasures against well known anti-malware products and may prevent them from running or being installed, or even uninstall them. An example of one that uses all three methods is Gromozon, a new breed of malware. It uses alternate data streams to hide. A rootkit hides it even from alternate data streams scanners and actively stops popular rootkit scanners from running.

Security practices

To detect spyware, computer users have found several practices useful in addition to installing anti-spyware programmes. Many system operators install a web browser other than IE, such as Opera, Google Chrome or Mozilla Firefox. Though no browser is completely safe, Internet Explorer is at a greater risk for spyware infection due to its large user base as well as vulnerabilities such as ActiveX.. Some ISPs—particularly colleges and universities—have taken a different approach to blocking spyware: they use

their network firewalls and web proxies to block access to Web sites known to install spyware.

On March 31, 2005, Cornell University's Information Technology department released a report detailing the behaviour of one particular piece of proxy-based spyware, *Marketscore*, and the steps the university took to intercept it. Many other educational institutions have taken similar steps. Spyware programmes which redirect network traffic cause greater technical-support problems than programmes which merely display ads or monitor users' behaviour, and so may more readily attract institutional attention.

Some users install a large hosts file which prevents the user's computer from connecting to known spyware-related web addresses. However, by connecting to the numeric IP address, rather than the domain name, spyware may bypass this sort of protection. Spyware may get installed via certain shareware programmes offered for download. Downloading programmes only from reputable sources can provide some protection from this source of attack. Recently, CNet revamped its download directory: it has stated that it will only keep files that pass inspection by Ad-Aware and Spyware Doctor.

The first step to removing spyware is to put a computer on "lockdown". This can be done in various ways, such as using anti-virus software or simply disconnecting the computer from the internet. Disconnecting the internet prevents controllers of the spyware from being able to remotely control or access the computer. The second step to removing the spyware is to locate it and remove it,

manually or through use of credible anti-spyware software. During and after lockdown, potentially threatening websites should be avoided.

Programmes Distributed with Spyware

- Bonzi Buddy
- Dope Wars
- EDonkey2000
- Grokster
- Kazaa
- Morpheus
- RadLight
- Sony's Extended Copy Protection involved the installation of spyware from audio compact discs through autorun. This practice sparked considerable controversy when it was discovered.
- WeatherBug
- WildTangent The antispyware programme Counterspy used to say that it's okay to keep WildTangent, but it now says that the spyware Winpipe is "possibly distributed with the adware bundler WildTangent or from a threat included in that bundler".

Programmes Formerly Distributed with Spyware

- AOL Instant Messenger (AOL Instant Messenger still packages Viewpoint Media Player, and WildTangent)
- DivX (except for the paid version, and the "standard" version without the encoder). DivX announced removal of GAIN software from version 5.2.

- FlashGet (trial version prior to programme being made freeware)
- magicJack

Rogue Anti-spyware Programmes

Malicious programmers have released a large number of rogue (fake) anti-spyware programmes, and widely distributed Web banner ads now spuriously warn users that their computers have been infected with spyware, directing them to purchase programmes which do not actually remove spyware—or else, may add more spyware of their own. The recent proliferation of fake or spoofed antivirus products has occasioned some concern. Such products often bill themselves as antispyware, antivirus, or registry cleaners, and sometimes feature popups prompting users to install them. This software is called rogue software. It is recommended that users do not install any freeware claiming to be anti-spyware unless it is verified to be legitimate. Some known offenders include:

- AntiVirus 360
- Antivirus 2008
- Antivirus 2009
- AntiVirus Gold
- ContraVirus
- MacSweeper
- Pest Trap
- PSGuard
- Spy Wiper
- Spydawn

- Spylocked
- Spysheriff
- SpyShredder
- Spyware Quake
- SpywareStrike
- UltimateCleaner
- WinAntiVirus Pro 2006
- Windows Police Pro
- WinFixer
- WorldAntiSpy

Fake antivirus products constitute 15 percent of all malware. On January 26, 2006, Microsoft and the Washington state attorney general filed suit against Secure Computer for its Spyware Cleaner product. On December 4, 2006, the Washington attorney general announced that Secure Computer had paid $1 million to settle with the state. As of that date, Microsoft's case against Secure Computer remained pending.